

---

# Analyzing text to extract information about gene-drug-disease interactions

Russ B. Altman, MD, PhD  
Stanford University

Rawls-Palmer Lecture  
March 16, 2017

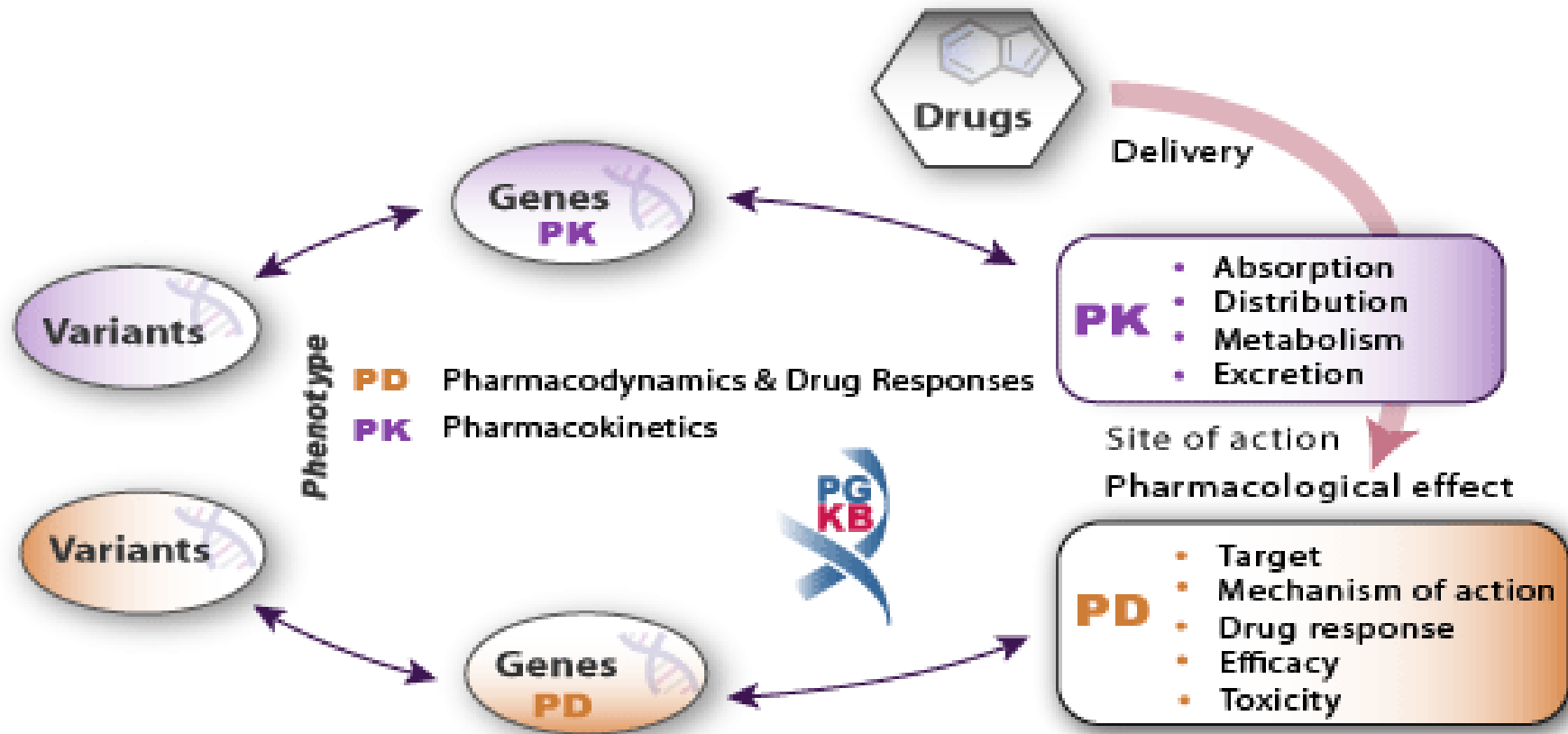
PharmGKB, <http://www.pharmgkb.org/>

---

# Thanks to colleagues at ASCPT.

- Kathy Giacomini
- Darrel Abernethy
- Neal Benowitz
- Terry Blaschke
- Dan Roden
- Don Stanski
- Dr. & Mrs. W.B. Rawls—who founded this award "to help bridge the gap between the results of research and its application in patient care."


# Variation in genes involved in drug metabolism and/or drug response



Search PharmGKB:

**What is the PharmGKB?**

*Find out how we go from extraction of gene-drug relationships in the literature to implementation of pharmacogenomics in the clinic...*



**Latest Blog Posts**

Associate Director of PharmGKB Discussing Database Curation and Genetic Test Interpretation at FDA Workshop

Concordance of Drug Labels and Clinical Annotations

# http://www.pharmgkb.org/

**Guidelines**

- Well-known PGx associations
- Clinically relevant PGx summaries
- PGx drug dosing guidelines
- Drug labels with PGx info
- PGx gene haplotypes

**CPIC Guidelines**

- See all CPIC guidelines
- Recent guidelines:
  - UGT1A1/atazanavir: [article](#) and [supplement](#)
  - SSRIs: citalopram/fluvoxamine/paroxetine [article](#) and [supplement](#)
- CPIC genes/drugs of interest
- TPP gene tables

**Resources**

- VIP:** Very Important PGx gene summaries
- PharmGKB pathways
- Annotated SNPs by gene
- Drugs with genetic information

**CPIC: Implementing PGx**  
a PharmGKB & PGRN collaboration

Follow us on:



Get your PGx fix:




PharmGKB is a partner of the



Pharmacogenomics Research Network


# PharmGKB on desktop


Welcome to the new PharmGKB! We'd love your feedback, [click here](#) to take the survey. ✕


 Publications News Downloads Contact Account


Search PharmGKB... Q

Search for a chemical, gene, variant, or combination

Drugs  609

Pathways  115

Dosing Guidelines  96

Drug Labels  449

**WHAT IS PHARMACOGENOMICS?**

The study of the relationship between genetic variations and how our body responds to medications.





[Pretty cool right? Tell me more...](#)


**PHARMACOGENOMICS. KNOWLEDGE. IMPLEMENTATION.**

PharmGKB is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers.

[Learn more about PharmGKB](#)

**Annotations**

Clinical	Research
 DOSING GUIDELINES 96	 PATHWAYS 115
 DRUG LABELS 449	 VIPs (Very Important Pharmacogenes) 64



# PharmGKB on tablet

5:54 PM  
next.pharmgkb.org

Welcome to the new PharmGKB! We'd love your feedback, [click here](#) to take the survey.

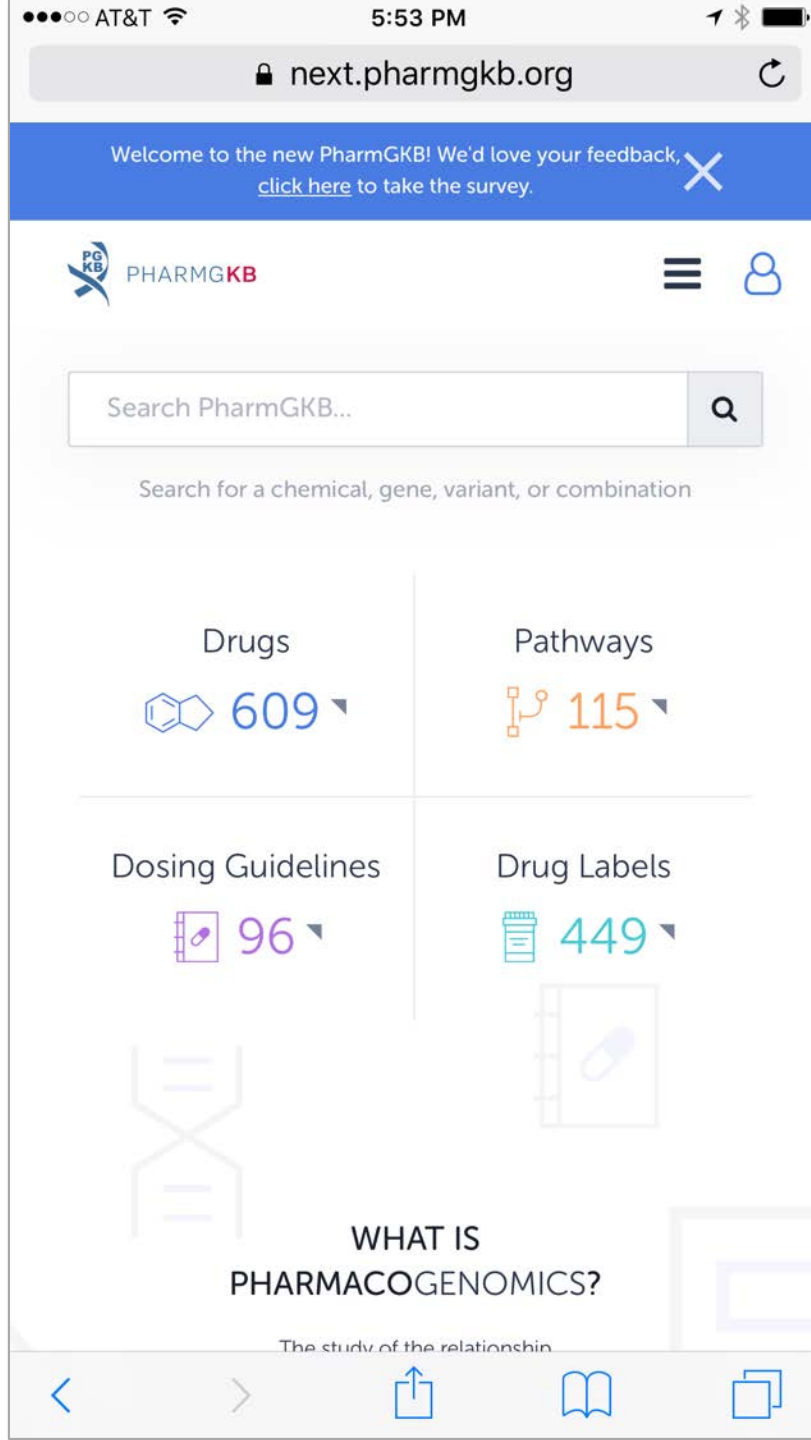
PHARMGKB

Publications News Downloads Contact Account

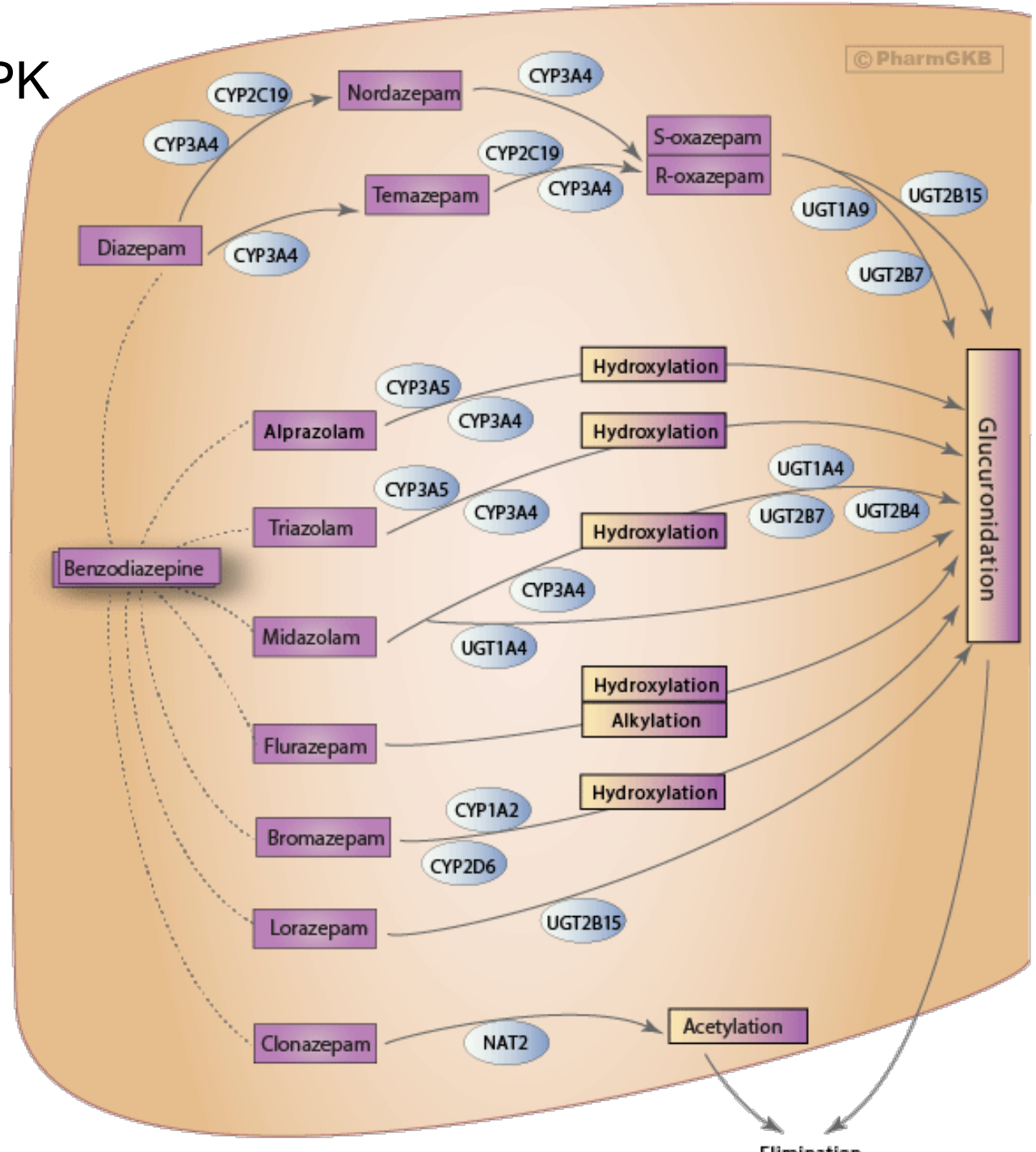
## Annotations

Clinical		Research		
	DOSING GUIDELINES	96	PATHWAYS	115
	DRUG LABELS	449	VIPs (Very Important Pharmacogenes)	64
	CLINICAL ANNOTATIONS	3,241	VARIANT ANNOTATIONS	17,306

# PharmGKB on phone

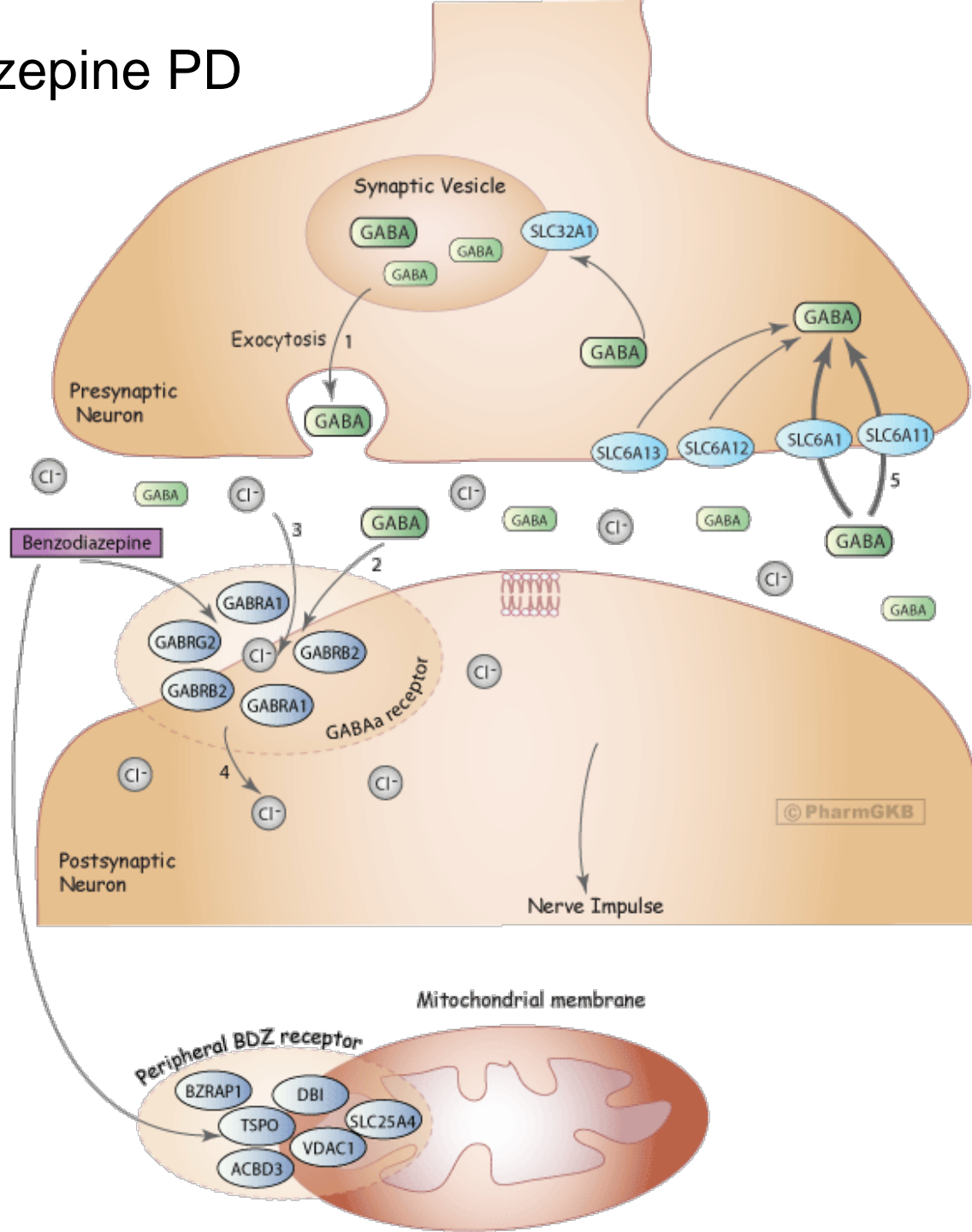


# Benzodiazepine PK





# Benzodiazepine PD



## Clinical Annotation for rs12248560 and clopidogrel

Level of Evidence 

**Level 1A**

### Type

Dosage, Efficacy,  
Toxicity/ADR

### Genes

[CYP2C19](#)

### Diseases

[Acute coronary syndrome](#),  
[Coronary Artery Disease](#),  
[Myocardial Infarction](#)

### OMB Race

Mixed Population

<b>CC</b>	Patients with the CC genotype (*1/*1): 1) m decreased, but not absent, risk for bleeding genotype 3) may have an increased risk for CT or TT genotype. Other genetic, including rs4986893, and clinical factors may also inf events.
<b>CT</b>	Patients with the CT (*1/*17) genotype: 1) n increased risk of bleeding with clopidogrel a decreased, but not absent, risk for adverse genotype. Current CPIC guidelines recomm as of yet. Other genetic, including CYP2C1: clinical factors may also influence a patient'
<b>TT</b>	Patients with the TT (*17/*17) genotype: 1) increased risk of bleeding with clopidogrel a decreased, but not absent, risk for adverse genotype. Current CPIC guidelines recomm genotype as of yet. Other genetic, including rs4986893, and clinical factors may also inf events.

 [View Evidence](#)

PMID: 20435227

# Clinical assessment incorporating a personal genome



*Evan A Ashley, Atul J Butte, Matthew T Wheeler, Rong Chen, Teri E Klein, Frederick E Dewey, Joel T Dudley, Kelly E Ormond, Aleksandra Pavlovic, Alexander A Morgan, Dmitry Pushkarev, Norma F Neff, Louanne Hudgins, Li Gong, Laura M Hodges, Dorit S Berlin, Caroline F Thorn, Katrin Sangkuhl, Joan M Hebert, Mark Woon, Hersh Sagreiya, Ryan Whaley, Joshua W Knowles, Michael F Chou, Joseph V Thakuria, Abraham M Rosenbaum, Alexander Wait Zaranek, George M Church, Henry T Greely, Stephen R Quake, Russ B Altman*

## Summary

**Background** The cost of genomic information has fallen steeply, but the clinical translation of genetic risk estimates remains unclear. We aimed to undertake an integrated analysis of a complete human genome in a clinical context.

**Methods** We assessed a patient with a family history of vascular disease and early sudden death. Clinical assessment included analysis of this patient's full genome sequence, risk prediction for coronary artery disease, screening for causes of sudden cardiac death, and genetic counselling. Genetic analysis included the development of novel methods for the integration of whole genome and clinical risk. Disease and risk analysis focused on prediction of genetic risk of variants associated with mendelian disease, recognised drug responses, and pathogenicity for novel variants. We queried disease-specific mutation databases and pharmacogenomics databases to identify genes and mutations with known associations with disease and drug response. We estimated post-test probabilities of disease by applying likelihood ratios derived from integration of multiple common variants to age-appropriate and sex-appropriate pre-test probabilities. We also accounted for gene-environment interactions and conditionally dependent risks.

**Findings** Analysis of 2.6 million single nucleotide polymorphisms and 752 copy number variations showed increased genetic risk for myocardial infarction, type 2 diabetes, and some cancers. We discovered rare variants in three genes

*Lancet* 2010; 375: 1525-35

See [Comment](#) page 1497

See [Online/Viewpoint](#)

DOI:10.1016/S0140-6736(10)60599-5

Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine (E A Ashley MRCP, M T Wheeler MD, F E Dewey MD, J W Knowles MD, A Pavlovic BS), Department of Medicine (Prof R B Altman MD), Department of Bioengineering (S R Quake PhD, D Pushkarev, N F Neff PhD, Prof R B Altman), Division of Medical Genetics (Prof L Hudgins MD).

# Summary of Pharmacogenetic Good News

Drug	Summary	Level of Evidence	PMID	Gene	rsID
HMG CoA Reductase Inhibitors (statins)	No increased risk of myopathy	High	18650507	SLCO1B1	rs4149056
Statins	No increased risk of myopathy	High	12811365	SLCO1B1	rs4149056
Desipramine; Fluoxetine	Depression may improve more than average	Medium	19414708	BDNF	rs61888800
Fluvastatin	Good response	Medium	18781850	SLCO1B1	rs11045819
Metoprolol and other CYP2D6 substrates	Normal CYP2D6 metabolizer.	Medium	19037197	CYP2D6	rs3892097/rs1800716
Pravastatin	May have good response	Medium	15199031	HMGCR	rs17238540
Pravastatin, Simvastatin	No reduced efficacy	Medium	15199031	HMGCR	rs17244841
Caffeine	No increased risk of heart problems with caffeine	Low	16522833	CYP1A2	rs762551
Calcium channel blockers	No increased risk of Torsades de Pointe	Low	15522280	KCNH2	rs36210421
Carbamazepine	SNP is part of protective haplotype for hypersensitivity to carbamazepine	Low	16538175	HSPA1A	rs1043620
Neviraprine	Reduced risk of hepatotoxicity	Low	16912957	ABCB1	rs1045642
Efavirenz; Nevirapine	Reduced risk of hepatotoxicity	Low	16912956	ABCB1	rs1045642
Epoetin Alfa	Lower dose of iron and epo required	Low	18025780	HFE	rs1799945
Fexofenadine	Average blood levels expected	Low	11503014	ABCB1	rs1045642
Irbesartan	Irbesartan may work better than beta-blocker	Low	15453913	APOB	rs1367117
Lithium	Increased likelihood of response	Low	18408563	CACNG2	rs5750285

# Summary of Pharmacogenetic Bad News

Drug	Summary	Level of Evidence	PMID	Gene	rsID
Clopidogrel & CYP2C19 substrates	CYP2C19 poor metabolizer, many drugs may need adjustment.	High	19106084	CYP2C19	rs4244285
Warfarin	Requires lower dose	High	15888487	VKORC1	rs9923231
Warfarin	Requires lower dose	High	19270263	CYP4F2	rs2108622
Metformin	Less likely to respond	Medium	18544707	CDKN2A/B	rs10811661
Troglitazone	Less likely to respond	Medium	18544707	CDKN2A/B	rs10811661
Cisplatin	Increased risk of nephrotoxicity	Low	19625999	SLC22A2	rs316019
Citalopram	May increase risk of suicidal ideation during therapy	Low	17898344	GRIA3	rs4825476
Escitalopram; Nortriptyline	Depression may not respond as well	Low	19365399	NR3C1	rs10482633
Morphine	May require higher dose for pain relief	Low	17156920	COMT	rs4680
Paclitaxel	Cancer may respond less well	Low	18836089	ABCB1	rs1045642
Pravastatin	May require higher dose	Low	15116054	SLCO1B1	rs2306283
Talinolol	May require higher dose	Low	18334920	ABCC2	rs2273697
Sildenafil	May not respond as well	Low	12576843	GNB3	rs5443



# = “CPIC”

## Clinical Pharmacogenetics Implementation Consortium Guidelines for *CYP2C9* and *VKORC1* Genotypes and Warfarin Dosing

JA Johnson<sup>1</sup>, L Gong<sup>2</sup>, M Whirl-Carrillo<sup>2</sup>, BF Gage<sup>3</sup>, SA Scott<sup>4</sup>, CM Stein<sup>5</sup>, JL Anderson<sup>6</sup>, SE Kimmel<sup>7,8,9</sup>, MTM Lee<sup>10</sup>, M Pirmohamed<sup>11</sup>, M Wadelius<sup>12</sup>, TE Klein<sup>2</sup> and RB Altman<sup>2,13</sup>

Warfarin is a widely used anticoagulant with a narrow therapeutic index and large interpatient variability in the dose required to achieve target anticoagulation. Common genetic variants in the cytochrome P450-2C9 (*CYP2C9*) and vitamin K-epoxide reductase complex (*VKORC1*) enzymes, in addition to known nongenetic factors, account for ~50% of warfarin dose variability. The purpose of this article is to assist in the interpretation and use of *CYP2C9* and *VKORC1* genotype data for estimating therapeutic warfarin dose to achieve an INR of 2–3, should genotype results be available to the clinician. The Clinical Pharmacogenetics Implementation Consortium (CPIC) of the National Institutes of Health Pharmacogenomics Research Network develops peer-reviewed gene–drug guidelines that are published and updated periodically on <http://www.pharmgkb.org> based on new developments in the field <sup>1</sup>

among the adverse events most frequently reported to the US Food and Drug Administration (FDA) and one of the most common reasons for emergency room visits.<sup>6</sup>

Warfarin is often dosed empirically: an initial dose is prescribed, typically followed by at least weekly measurement of the INR and subsequent dose adjustment. The initial dose is often based on population averages (e.g., 3–5 mg/day), but stable doses to achieve an INR of 2–3 can range from 1–20 mg/day. The iterative process to define the appropriate dose can take weeks to months, and during this period patients are at increased risk of over- or under-anticoagulation and thus at risk of thromboembolism or bleeding.

### Warfarin pharmacology and pharmacokinetics

**Figure 1** highlights key elements of warfarin pharmacology and pharmacokinetics. Warfarin inhibits vitamin K-epoxide reduct-

# Augmenting the network of molecular and cellular data with textual information.

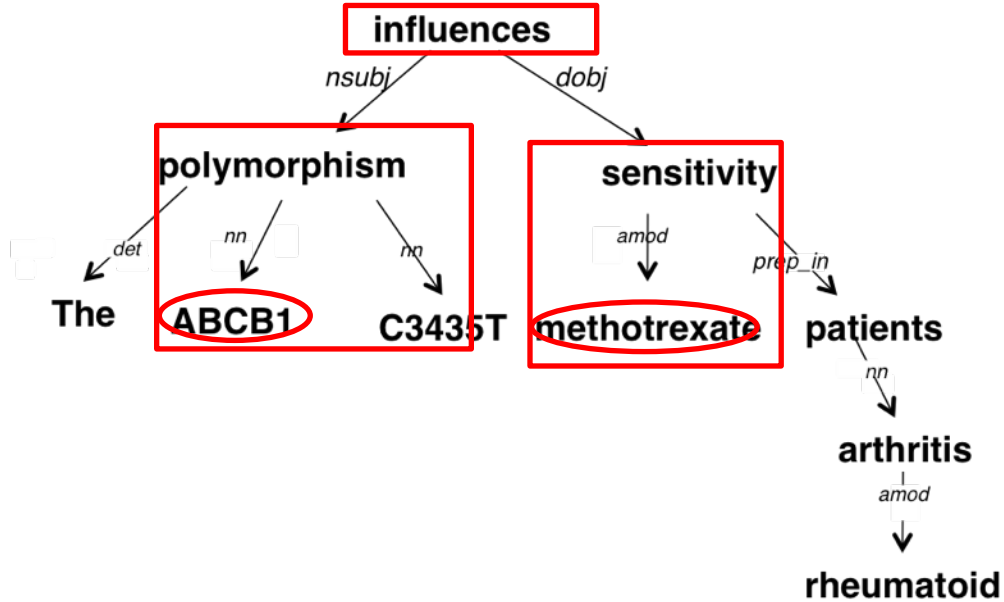
- PubMed now holds more than 24 million entries, most with abstracts
- Our biomedical knowledge is stored in the published literature

Can we have computers “read” PubMed abstracts and reason over them to predict new drug-drug interactions?

# Advances in natural language parsing enable high fidelity extraction of relations

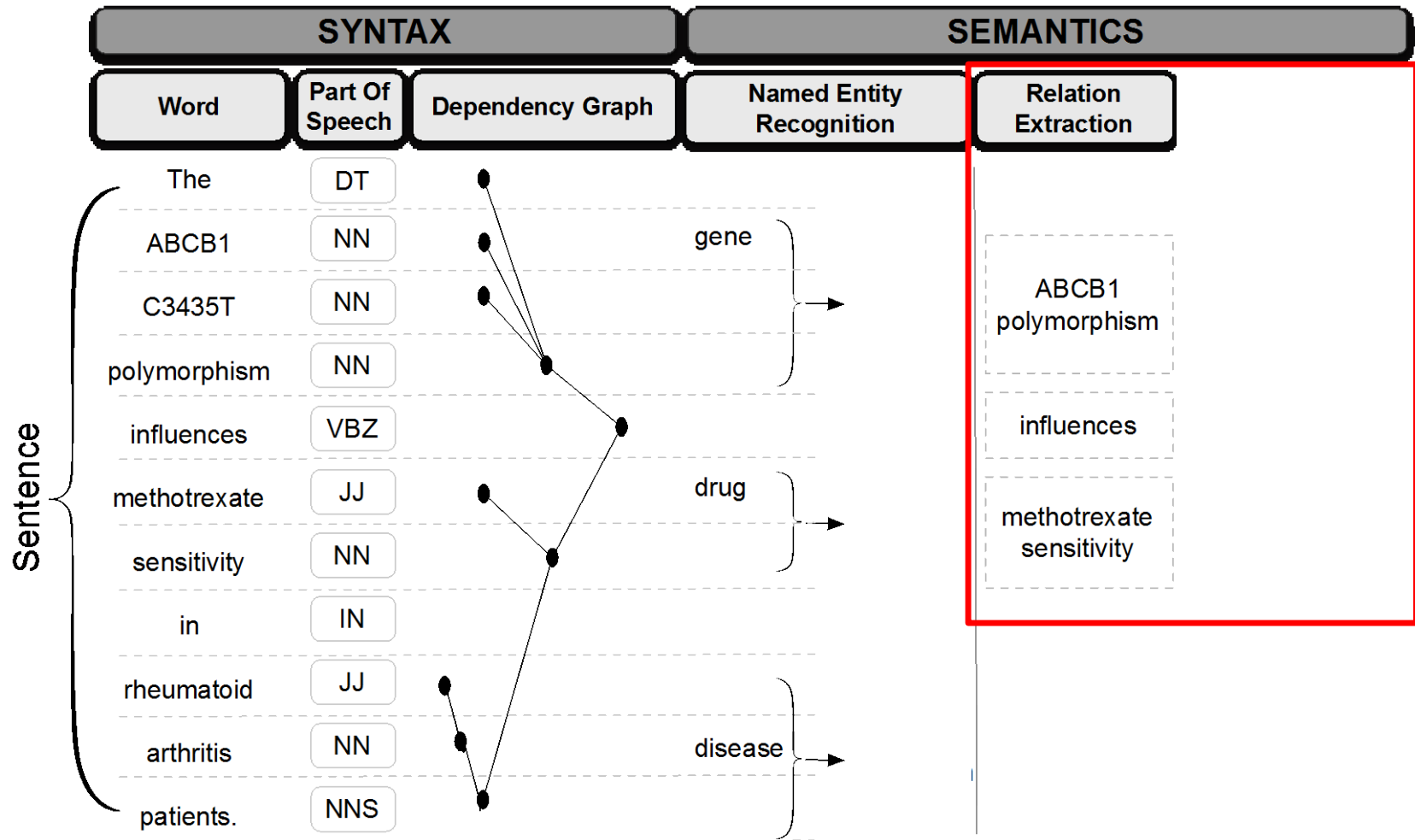
SYNTAX		
Word	Part Of Speech	Dependency Graph
The	DT	
ABCB1	NN	
C3435T	NN	
polymorphism	NN	
influences	VBZ	
methotrexate	JJ	
sensitivity	NN	
in	IN	
rheumatoid	JJ	
arthritid	NN	
patients.	NNS	

Dependency graph

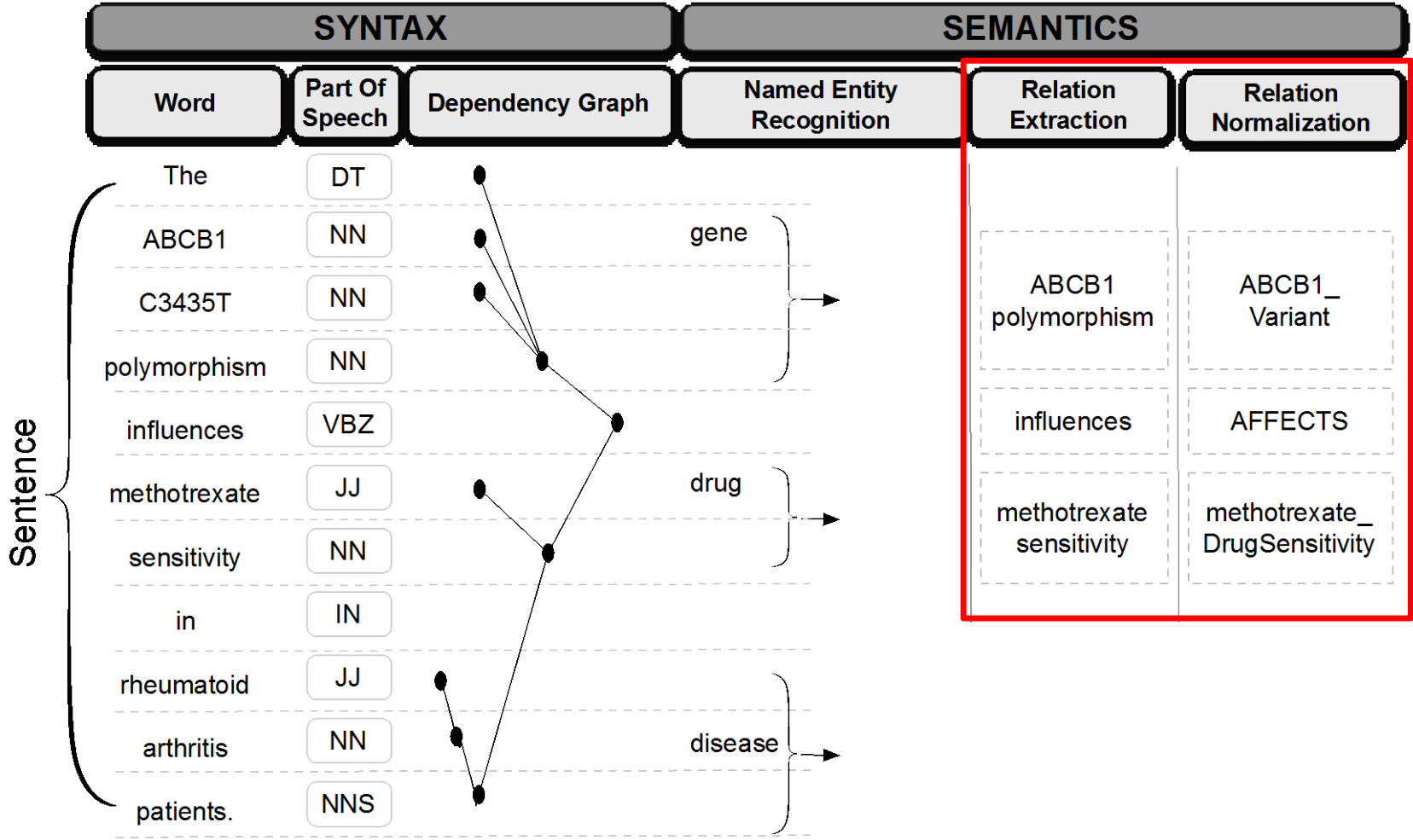




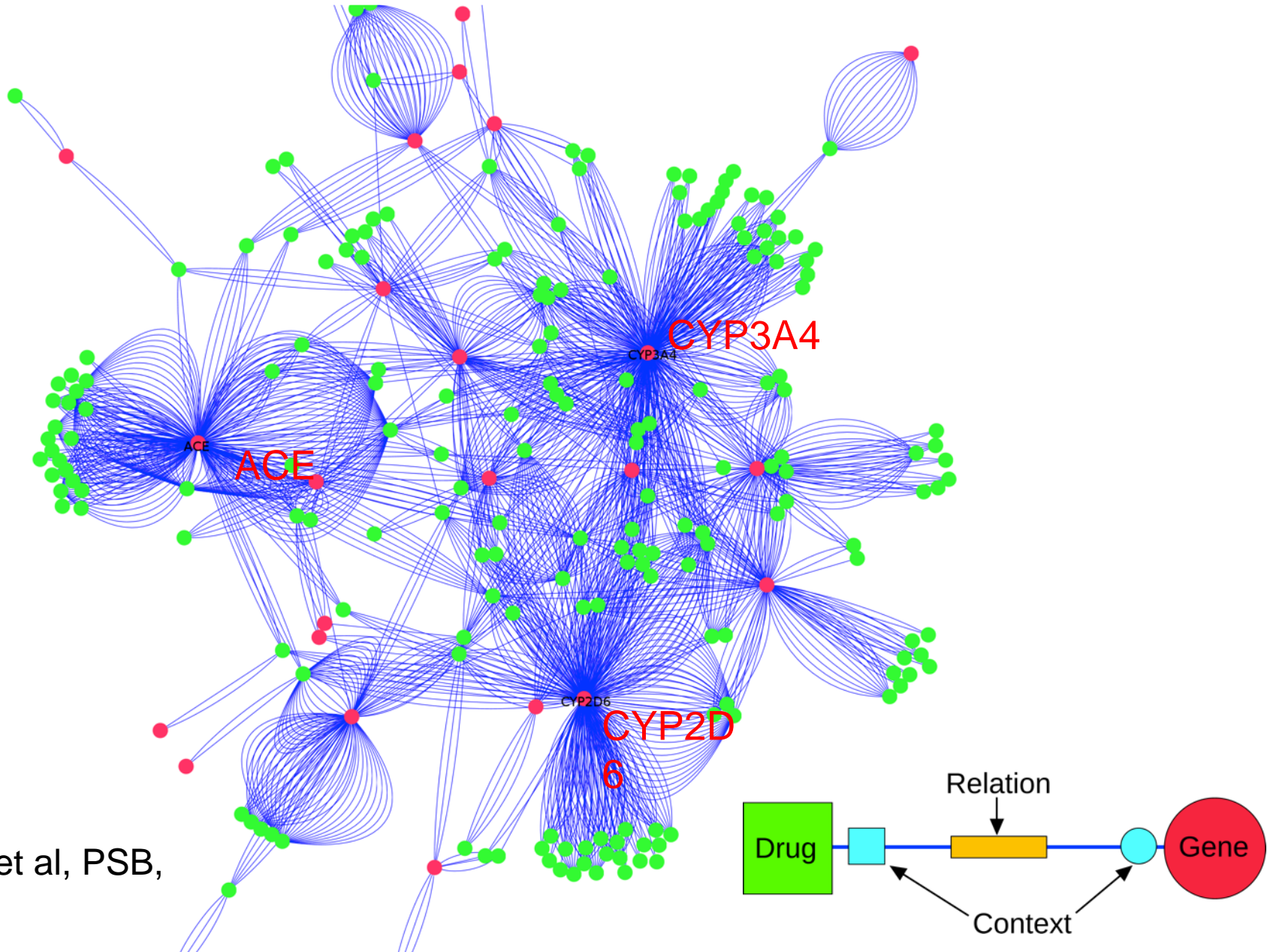
# We can identify genes, drugs, phenotypes in text using these technologies.



# KEY: we map extracted entities to standard terminologies

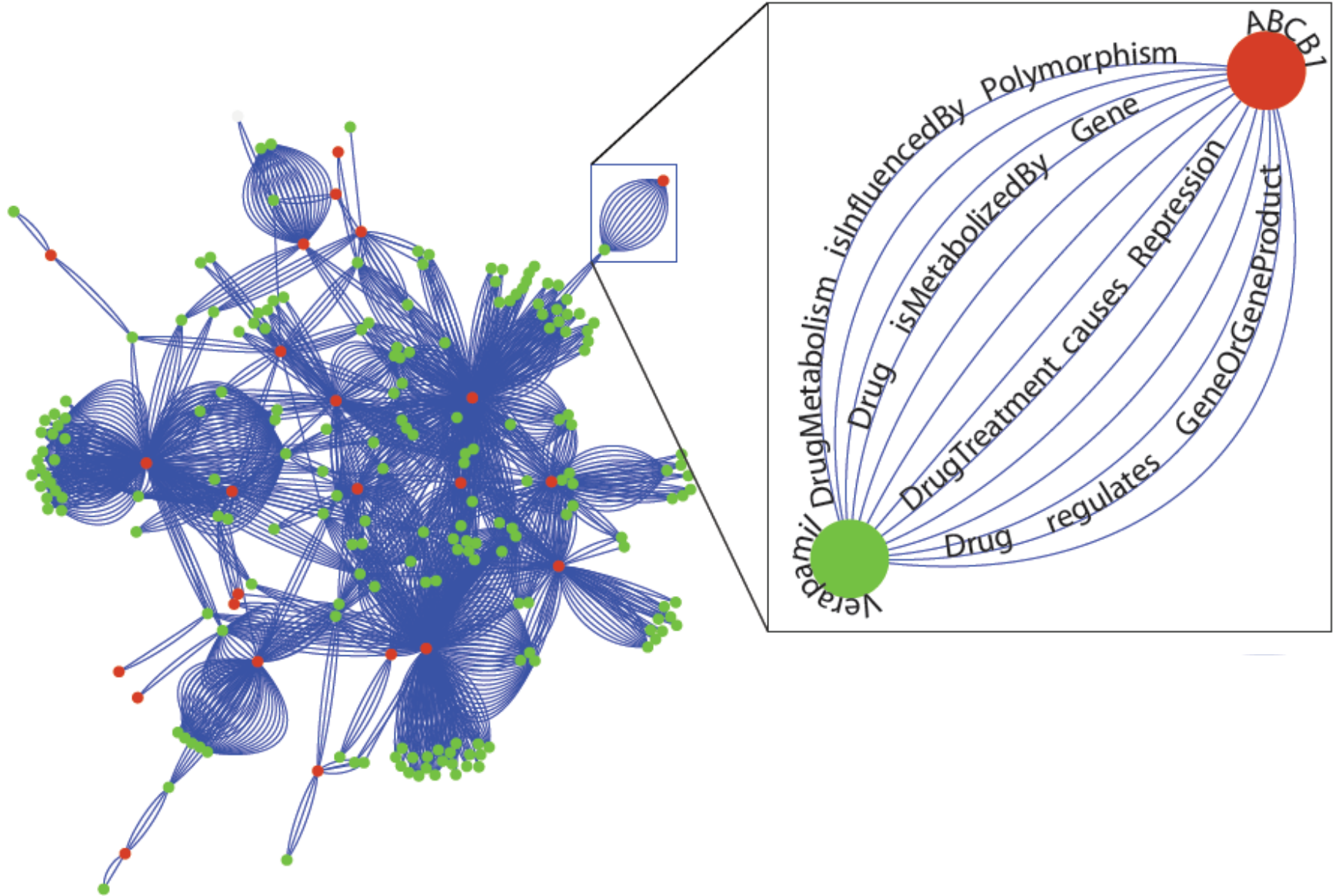


# Semantic network of 170,598 normalized relations from all PubMed abstracts involving 40 key genes.

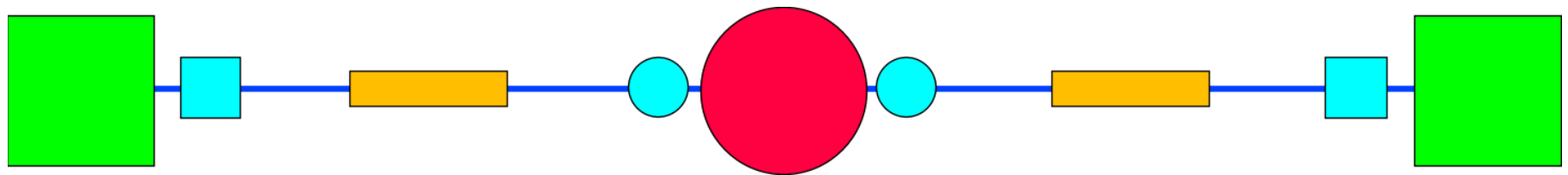
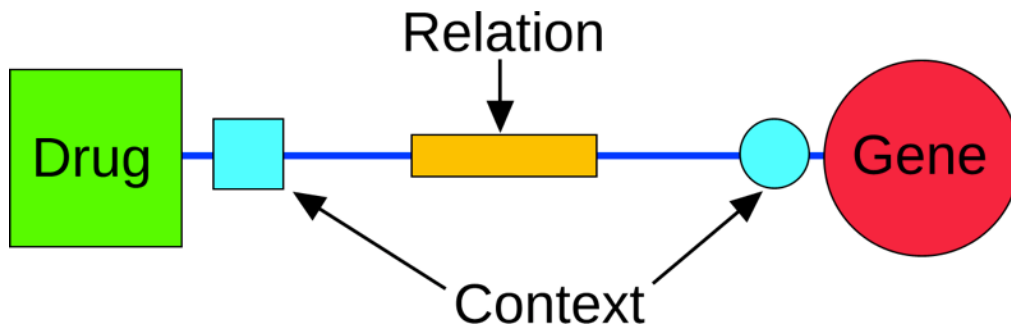


Percha et al, PSB,  
2012.

# ABCB1 gene and verapamil drug



We chain together two gene-drug relationships to create a drug-gene-drug relationship = potential drug interaction!



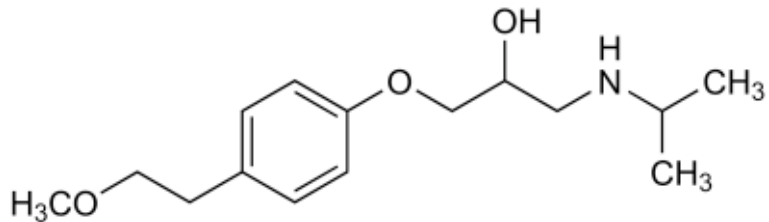
e.g. “Drug A decreases levels of Gene, but Gene metabolizes drug B.”

# Top predicted DDIs from text processing after training on known DDIs

drug pair	interact	npaths	avgvotes	p.glm	LAB
fluoxetine-diazepam	1*	975	59.68%	0.608	
tramadol-dextromethorphan	1	536	95.27%	0.533	
venlafaxine-dextromethorphan	1	528	94.99%	0.523	
naproxen-diazepam	0	806	65.60%	0.506	1
paroxetine-dextromethorphan	1	489	95.32%	0.489	
metoprolol-dextromethorphan	0	490	90.17%	0.441	0
lisinopril-enalapril	1--	986	37.46%	0.410	
verapamil-omeprazole	0	469	87.30%	0.395	1
dextromethorphan-codeine	1-	440	86.98%	0.367	
omeprazole-naproxen	1*	911	40.09%	0.366	
sertraline-dextromethorphan	1**	357	95.02%	0.365	
omeprazole-diazepam	1	872	38.84%	0.322	
verapamil-fluconazole	1	252	97.34%	0.296	
verapamil-fexofenadine	0	248	93.28%	0.262	1
verapamil-atorvastatin	1	296	88.32%	0.261	
naproxen-fluoxetine	1*	626	52.50%	0.240	
warfarin-glibenclamide	1*	221	92.29%	0.236	
warfarin-fluoxetine	1	272	86.93%	0.234	
verapamil-carvedilol	1	200	92.99%	0.227	
venlafaxine-tramadol	1	145	98.45%	0.227	
verapamil-clopidogrel	1-	184	94.40%	0.226	1
venlafaxine-paroxetine	1	132	99.18%	0.223	
verapamil-simvastatin	1	263	86.04%	0.222	
tramadol-paroxetine	1	133	98.42%	0.219	
warfarin-ibuprofen	1	172	94.41%	0.218	
omeprazole-dextromethorphan	0	150	96.36%	0.216	0
fluoxetine-dextromethorphan	1	160	95.16%	0.215	
venlafaxine-metoprolol	1*	132	94.89%	0.196	
venlafaxine-sertraline	1	97	97.97%	0.194	
tramadol-sertraline	1	97	97.84%	0.193	
dextromethorphan-citalopram	1**	104	96.35%	0.188	
paroxetine-metoprolol	1	126	93.79%	0.186	
tramadol-metoprolol	0	132	93.13%	0.186	0
verapamil-diazepam	1-	300	76.30%	0.185	1
dextromethorphan-aripiprazole	0	132	92.67%	0.183	0

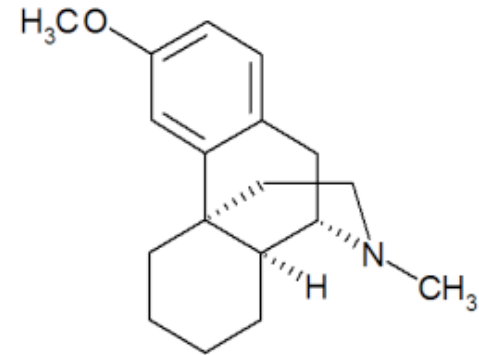


# Metoprolol



<b>Class</b>	Beta blocker
<b>Brand</b>	Lopressor, Toprol
<b>Treats</b>	High blood pressure Angina (chest pain) Heart attack (improves survival) Heart failure
<b>Effects</b>	Relaxes blood vessels Slows heart rate
<b>Mech</b>	Blocks beta receptors on sympathetic nerves

# Dextromethorphan



<b>Class</b>	Non-opioid antitussive
<b>Brand</b>	Robitussin (+ tons of others)
<b>Treats</b>	Cough
<b>Effects</b>	Acts on central nervous system to elevate threshold for coughing
<b>Mech</b>	NMDA and glutamate antagonist Blocks dopamine reuptake site (?)

# Metoprolol

# Dextromethorphan

## metoprolol Drug is Metabolized By Gene CYP2D6...

Drug Metab Dispos. 2000 Nov;28(11):1327-34.

### **Automated definition of the enzymology of drug oxidation by the major human drug metabolizing cytochrome P450s.**

McGinnity DF, Parker AJ, Soars M, Riley RJ.

Department of Physical & Metabolic Science, AstraZeneca R&D Charnwood, Loughborough, Leicestershire, United Kingdom.

#### **Abstract**

A fully automated assay to determine the enzymology of drug oxidation by the major human hepatic cytochrome P450s (CYPs; CYP1A2, -2C9, -2C19, -2D6, and -3A4) coexpressed functionally in *Escherichia coli* with human NADPH-P450 reductase has been developed and validated. Ten prototypic substrates were chosen for which clearance was primarily CYP-dependent, and the activities of these five major CYPs were represented. A range of intrinsic clearance (CL(int)) values were obtained for substrates in both pooled human liver microsomes (HLM; 1-380 microl. min<sup>-1</sup>mg<sup>-1</sup>) and recombinant CYPs (0.03-7 microl. min<sup>-1</sup>pmol<sup>-1</sup>) and thus the percentage contribution of individual CYPs toward their oxidative metabolism could be estimated. All the assignments were consistent with the available literature data. Tolbutamide was metabolized by CYP2C9 (70%) and CYP2C19 (30%), diazepam by CYP2C19 (100%),

ibuprofen by CYP2C9 (90%) and CYP2C19 (10%), and omeprazole by CYP2C19 (68%) and CYP3A4 (32%). Metoprolol and dextromethorphan were primarily CYP2D6 substrates and propranolol was metabolized by CYP2D6 (59%), CYP1A2 (26%), and CYP2C19 (15%). Diltiazem, testosterone, and verapamil were metabolized predominantly by CYP3A4. In addition, the metabolite profile for the CYP-dependent clearance of several markers determined by mass spectroscopy was as predicted from the literature. There was a good correlation between the sum of individual CYP CL(int) and HLM CL(int) ( $r^2 = 0.8$ ,  $P < .001$ ) for the substrates indicating that recombinant CYPs may be used to predict HLM CL(int) data. This report demonstrates that recombinant human CYPs may be useful as an approach for the prediction of the enzymology of human CYP metabolism early in the drug discovery process.



# Metoprolol

# Dextromethorphan

## ... CYP2D6 Gene metabolizes Drug Dextromethorphan

Can J Anaesth. 2005 Oct;52(8):806-21.

### [Genetic polymorphism and drug interactions: their importance in the treatment of pain].

[Article in French]

Samer CF, Piquet V, Dayer P, Desmeules JA.

Service de pharmacologie et toxicologie cliniques et Centre multidisciplinaire d'étude et de traitement de la douleur, Hôpitaux Universitaires de Genève, Genève, Suisse. Caroline.Samer@hcuge.ch

#### Abstract

**OBJECTIVES:** To evaluate the impact of certain genetic polymorphisms on variable responses to analgesics

**SOURCES:** Systematic review, by means of a structured computerized search in the Medline database (1966-2004). Articles in English and French were selected. References in relevant articles were also retrieved.

**MAIN FINDINGS:** Most analgesics are metabolized by CYP isoenzymes subject to genetic polymorphism. NSAIDs are metabolized by CYP2C9; opioids described as "weak" (codeine, tramadol), anti-depressants and dextromethorphan are metabolized by CYP2D6 and some "potent" opioids (buprenorphine, methadone or fentanyl) by CYP3A4/5. After the usual doses have been administered, drug toxicity or, on the contrary, therapeutic ineffectiveness may occur, depending on polymorphism and the substance. Drug interactions mimicking genetic defects because of the existence of CYP inhibitors and inducers, also contribute to the variable response to analgesics. Some opioids are substrates of P-gp, a transmembrane transporter also subject to genetic polymorphism. However, P-gp could only play a minor modulating role in man on the central effects of morphine, methadone and fentanyl.

**CONCLUSION:** In the near future, pharmacogenetics should enable us to optimize therapeutics by individualizing our approach to analgesic drugs and making numerous analgesics safer and more effective. The clinical usefulness of these individualized approaches will have to be demonstrated by appropriate pharmacoeconomic studies and analyses.

# Case Report: 2011

*Ann Pharmacother.* 2011 Jan;45(1):e1. doi: 10.1345/aph.1P301. Epub 2010 Dec 14.

## Myoclonus after dextromethorphan administration in peritoneal dialysis.

Tanaka A<sup>1</sup>, Nagamatsu T, Yamaguchi M, Nomura A, Nagura F, Maeda K, Tomino T, Watanabe T, Shimizu H, Fujita Y, Ito Y.

### Author information

#### Abstract

**OBJECTIVE:** To report a case of myoclonus that developed after administration of dextromethorphan.

**CASE SUMMARY:** A 64-year-old man was diagnosed with chronic renal failure due to diabetic nephropathy. The patient started on peritoneal dialysis 6 months before he was hospitalized. Two days before hospitalization, he developed cough and sputum and he visited an outpatient clinic, where dextromethorphan was prescribed. After taking a total of 30 mg of dextromethorphan, the patient developed myoclonus, tremor, agitation, slurred speech, and diaphoresis, which continued after he stopped taking the prescribed medicine. He visited an emergency department and was hospitalized for examination and treatment of myoclonus.

**DISCUSSION:** As the patient's dialysis schedule was adequate, these symptoms were likely not due to uremia. The blood concentration of dextromethorphan (2.68 ng/mL) 60 hours after the 30-mg dose was higher than expected, and the blood concentration of dextrorphan, a metabolite, was lower than expected. We suspected that myoclonus was due to dextromethorphan-related symptoms induced by CYP2D6, which primarily metabolizes dextromethorphan. We analyzed the CYP2D6 gene for polymorphisms and identified CYP2D6 (\*<sup>1</sup>/\*<sup>10</sup>). The patient had been taking metoprolol 40 mg/day for 2 years. The blood concentration of metoprolol 6 hours after administration was 13 ng/mL, which suggests that it was metabolized normally. Metoprolol has another metabolic pathway, via CYP2C19, and this may have led to its lack of accumulation. Moreover, metoprolol may have bound to active CYP2D6. Thus, affinity for CYP2D6, protein-binding rate, and lipid solubility may influence these drug interactions. Total scores for the Adverse Drug Reaction (ADR) probability scale and the Drug Interaction Probability Scale (DIPS) were 9 (highly probable) and 3 (possible), respectively.

**CONCLUSIONS:** Myoclonus and other symptoms in this patient may have been caused by a prolonged high concentration of dextromethorphan due to CYP2D6 polymorphisms and drug interactions.

PMID: 21228393 [PubMed - indexed for MEDLINE]

# Using textual context to define synonyms (with no training)

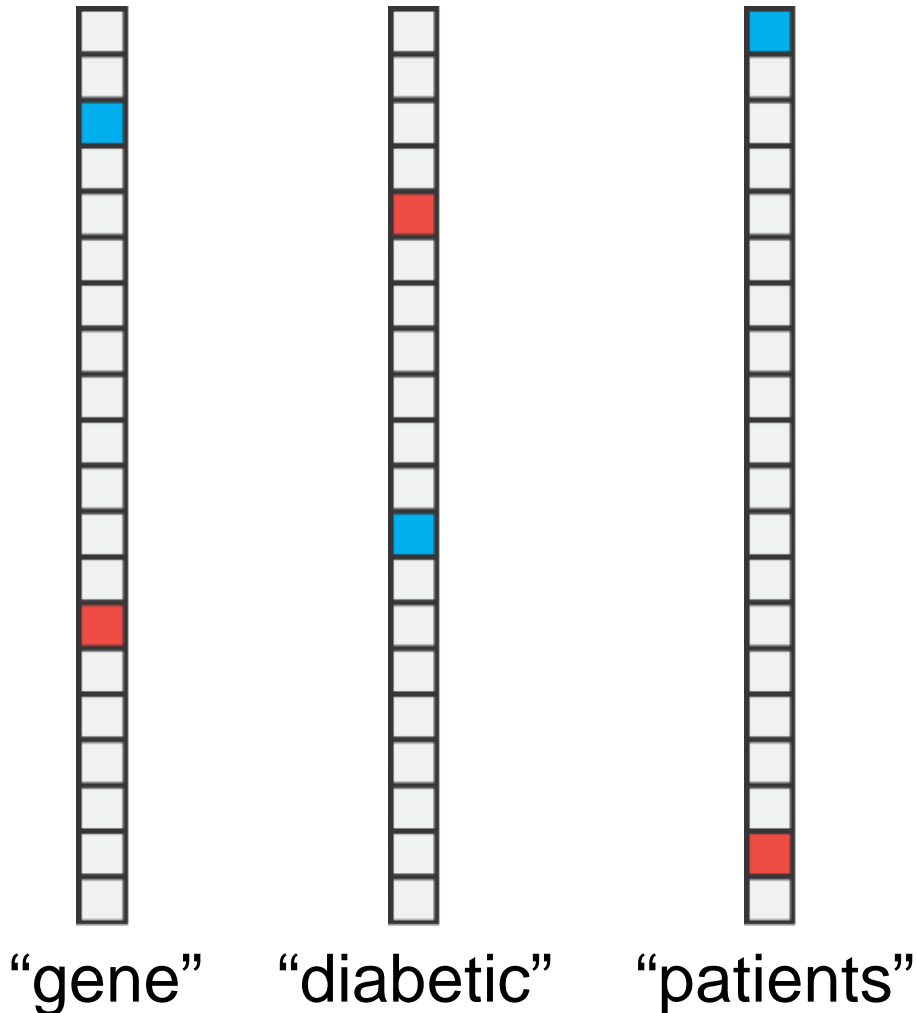
- “Random indexing”
- Each word gets unique random vector
- Context encoded by adding vectors of all surrounding words together = “context signature”
- Similarity of context measured by cosine of angle between context signature vectors.

Can we automatically learn synonyms for entities using this data-driven method?

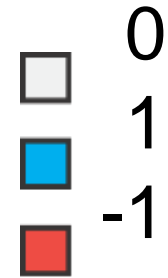
# What does “polymorphism” mean?

Our findings suggest that the myeloperoxidase G-463A polymorphism is a host genetic factor which determines the clinical outcome of *Helicobacter pylori* infection... Insertion / deletion polymorphism in the promoter of NFKB1 as a potential molecular marker for the risk of recurrence in superficial bladder cancer.... Genetic polymorphism and resistance mutations of HIV type 2 in antiretroviral-naive patients in Burkina Faso... One of these groups was representative of the genetic diversity previously found within the pathogen by random fragment length polymorphism and amplified fragment length polymorphism analyses... To identify mutations in mtDNA D-loop, polymerase chain reaction (PCR)-single strand conformation polymorphism (SSCP) analysis, followed by nucleotide sequencing, was performed... Their distinct sets of mtDNA polymorphism belonged to Eastern Asian haplogroup C4a1, while other previously identified six Chinese mitochondrial genomes... In the present study, we investigated the GSTM1 gene polymorphism in diabetic patients and healthy individuals and searched whether polymorphisms in GST genes are associated with diabetes mellitus (DM) in the Turkish population... on blood samples from 137 colorectal cancer patients and 199 controls using polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP)... Morphological changes of hippocampus, polymorphism of serotonin transporter gene, and down regulation of neurotrophin are also discussed in this review... fluorescent amplified fragment length polymorphism (FAFLP), enterobacterial repetitive intergenic consensus (ERIC) based genotyping and candidate orthologues sequencing revealed that MIP has been the predecessor of highly pathogenic *Mycobacterium* a

# Random indexing: an unsupervised method for assessing word similarity

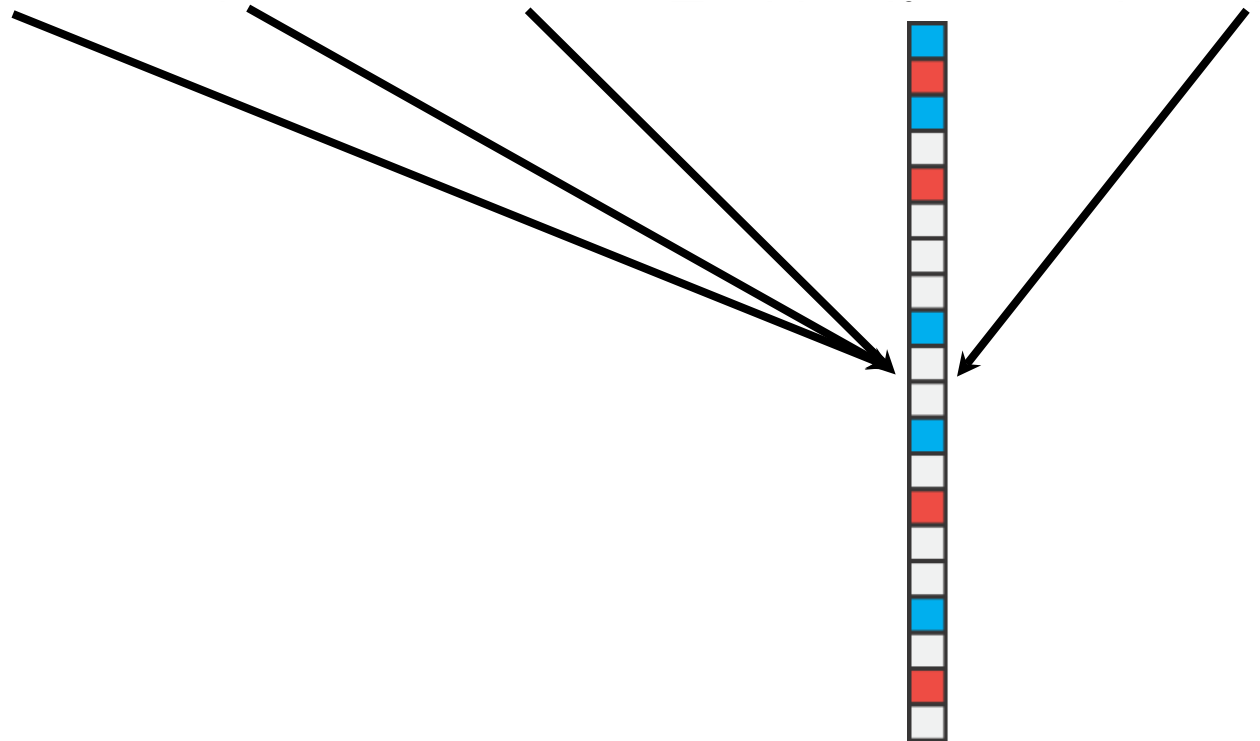


First, every word is assigned a randomly generated elemental vector.



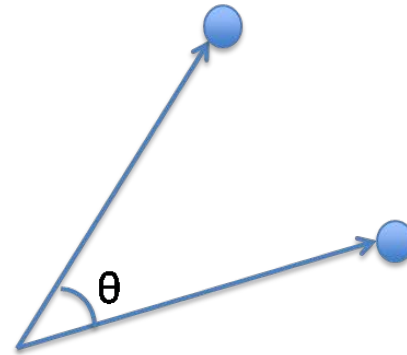
Context vectors are created by summing elemental vectors within a certain window

pathogen by random fragment length polymorphism and amp



Context vectors can be compared using cosine similarity to assess word similarity

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



# Specifics of our method: dataset and context vectors

- Corpus: 494,804 drug-gene sentences from Medline 2013 (real sentences of interest)
- Semantic Vectors implementation of random indexing (Java-based)
- Varied window size, vector dimension, seed length, word order encoding



# Recognizing synonyms with shared context

**Table 1.** Examples of high-ranking pairs of *concept* labels from drug-gene sentences, ordered by cosine similarity.

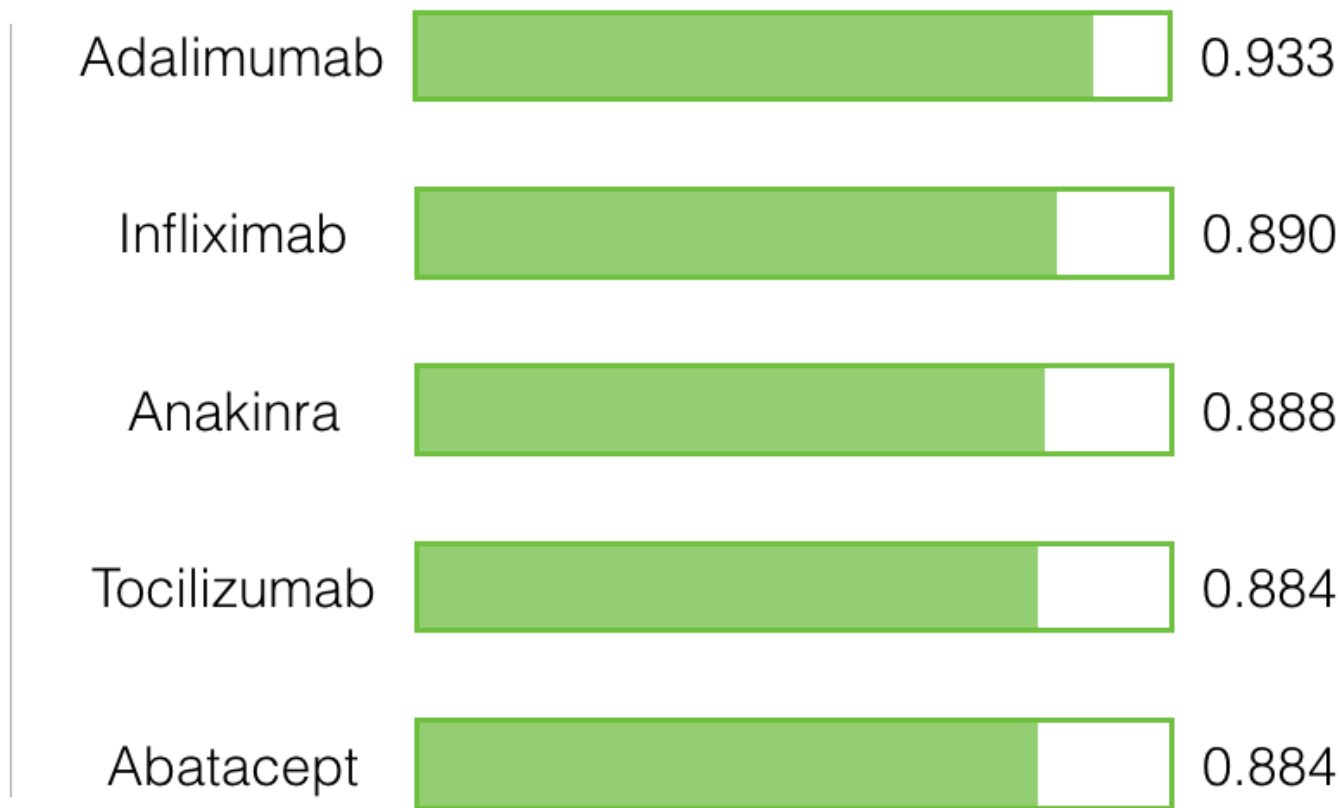
Concept Label 1	Concept Label 2	Cosine Similarity Score
inhibition	suppression	0.983
downregulation	upregulation	0.982
incidence	prevalence	0.981
assessment	evaluation	0.977
pharmacokinetics	disposition	0.974
association	interaction	0.973
inactivation	inhibition	0.973
tolerability	safety	0.972

**Table 2.** Examples of high-ranking pairs of *role* labels from drug-gene sentences, ordered by cosine similarity.

Role Label 1	Role Label 2	Cosine Similarity Score
investigate	examine	0.999
assess	evaluate	0.999
suggest	indicate	0.997
alter	affect	0.996
modulation	inhibition	0.992
suppress	stimulate	0.990
inhibit	prevent	0.988
catalyzed	catalysed	0.986

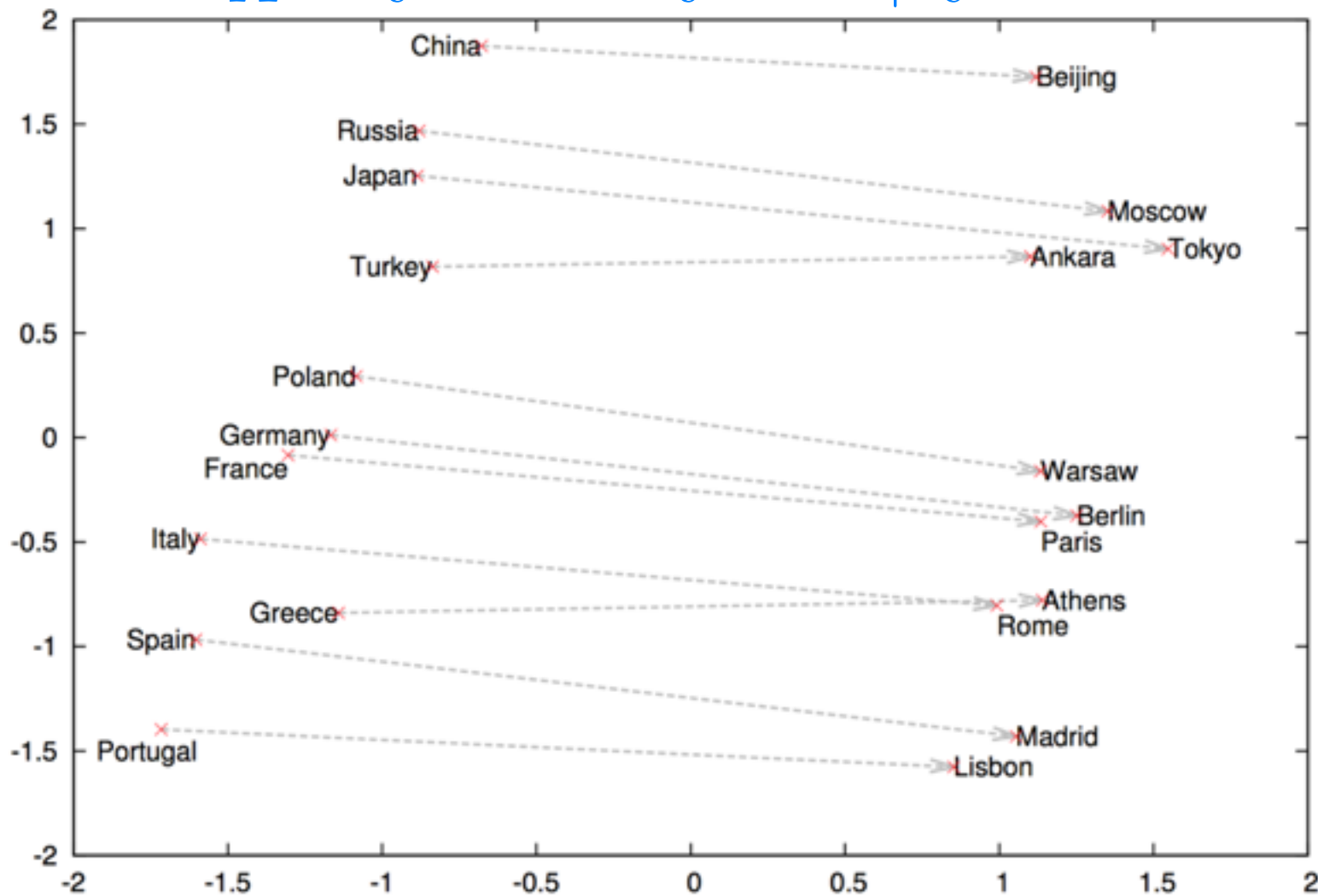
# What are most similar drugs? (using similar approach, word2vec)

Etanercept

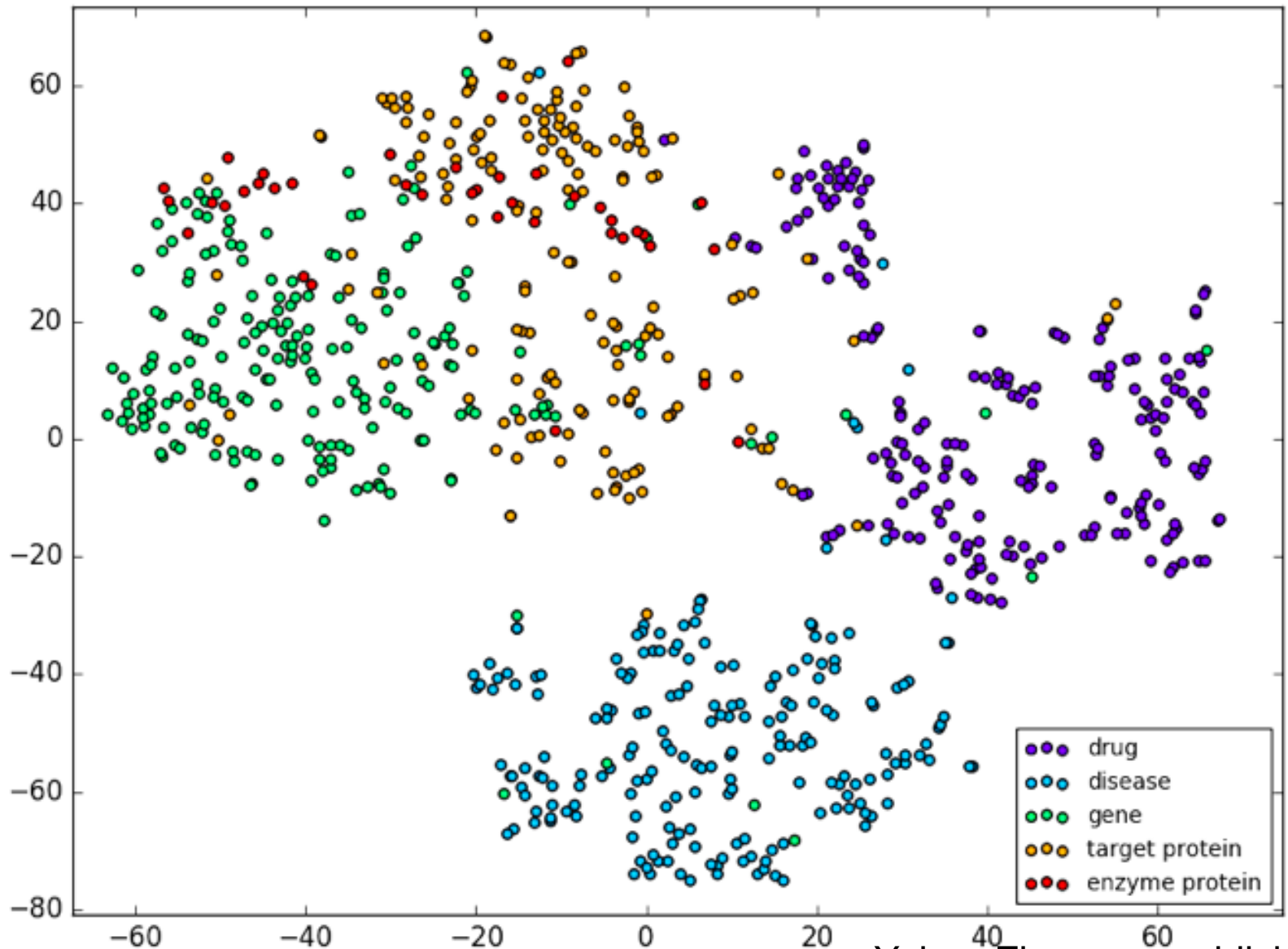


# Vector relationships preserve semantics!

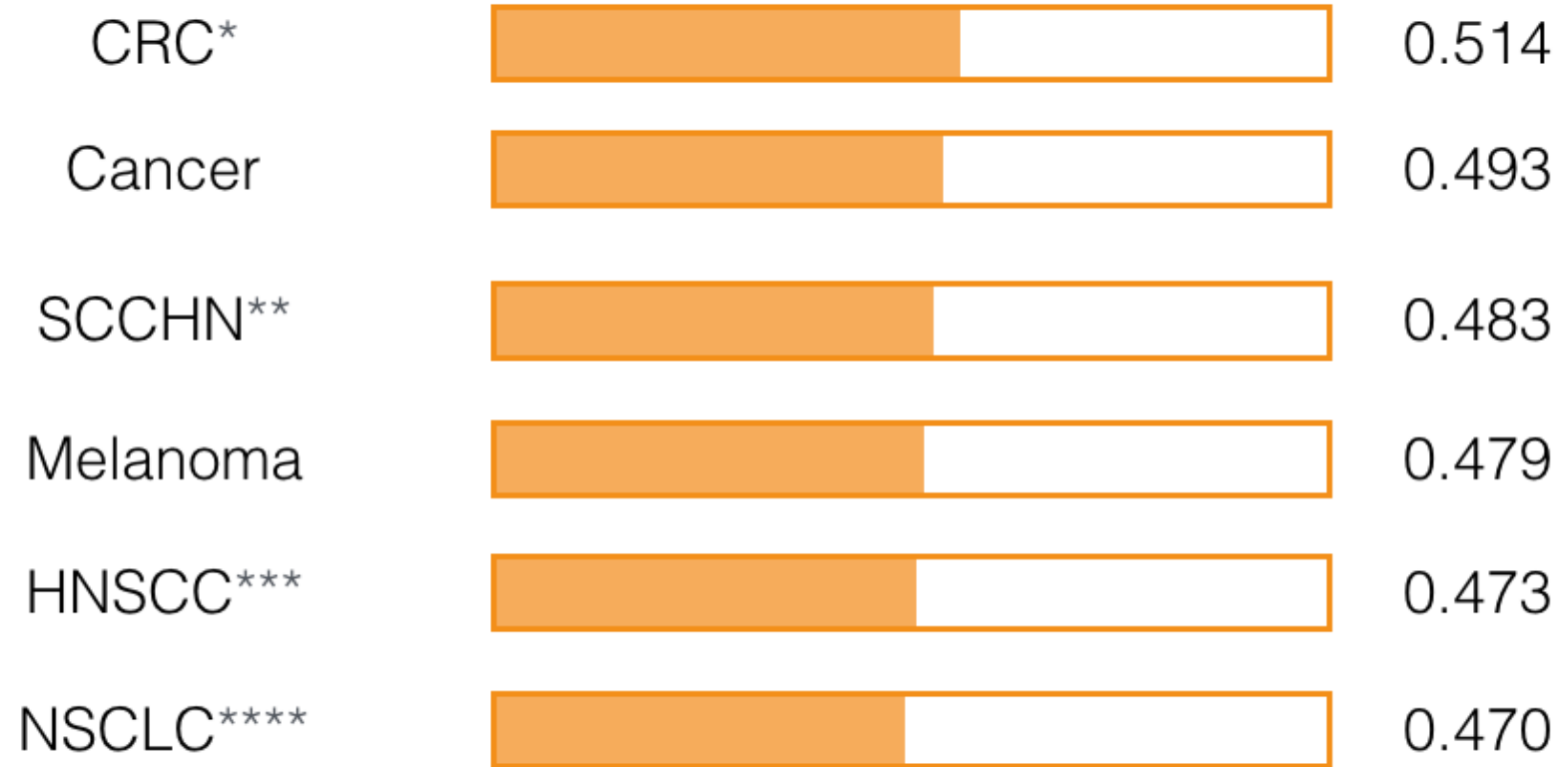
$$X = v^{\text{Beijing}} - v^{\text{China}} + v^{\text{Japan}}$$



# Similar analysis for drugs, diseases, genes (targets & enzymes). 24M abstracts. 2M words.



$$v^{\text{Etanercept}} - v^{\text{Arthritis}} + v^{\text{Cetuximab}} \approx$$



\* Colorectal Cancer

\*\* Squamous Cell Cancer of the Head & Neck

\*\*\* Head & Neck Cancer

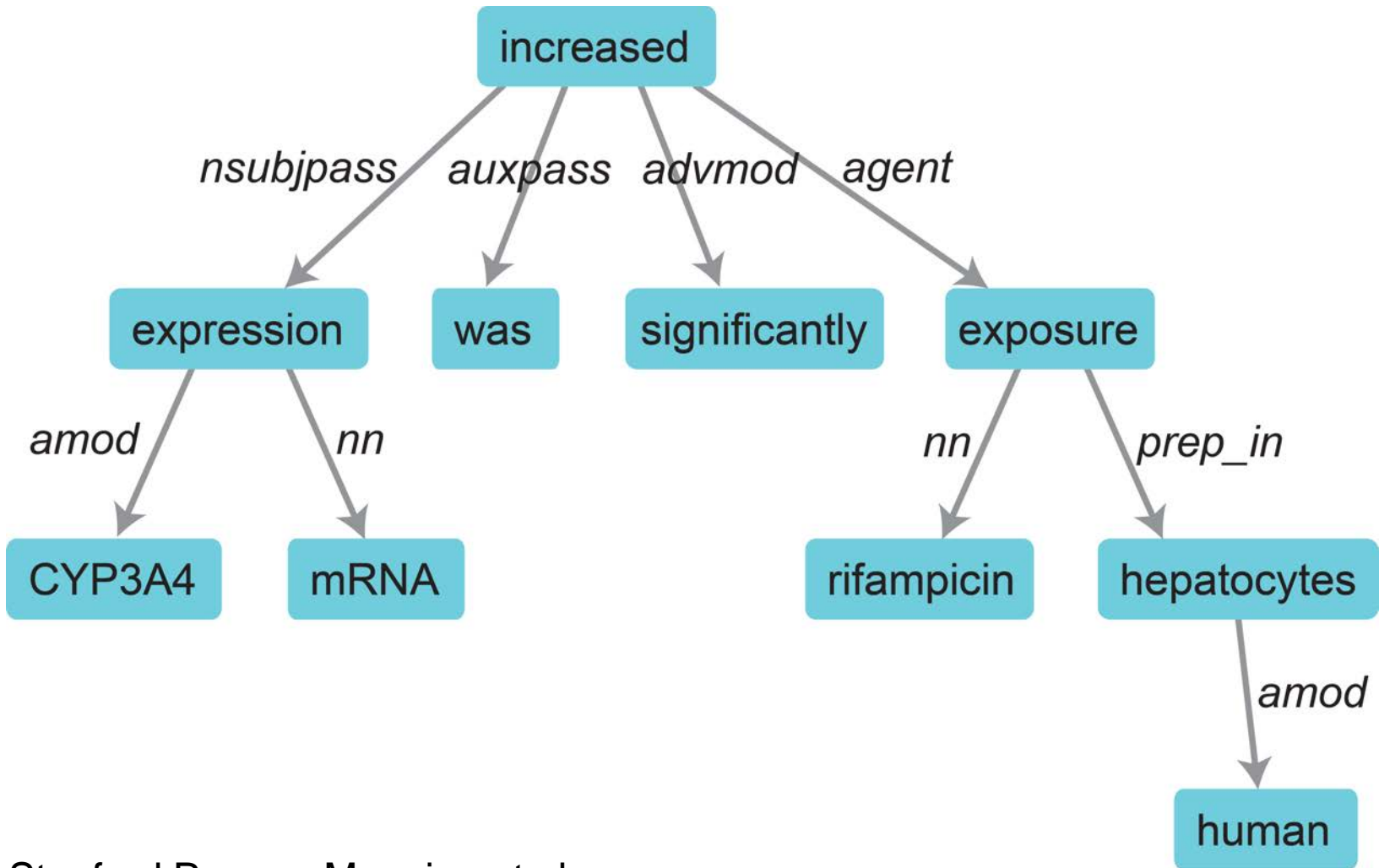
\*\*\*\* Non-small-cell Lung Carcinoma

# Characterizing the universe of gene-drug interactions.

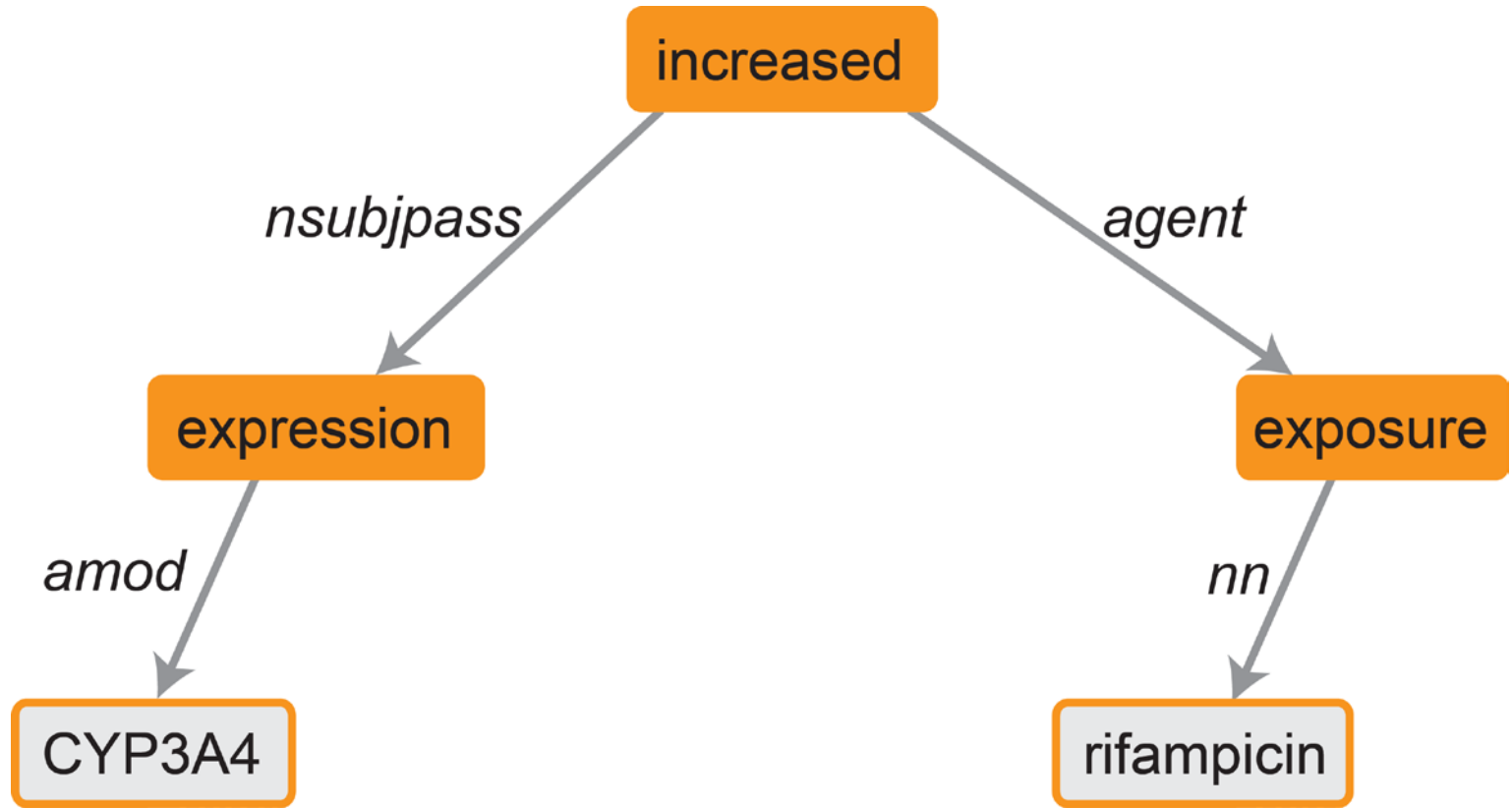
- More than  $20 \times 10^6$  PubMed abstracts
- Contain many assertions about relationships between genes/drugs—indeed perhaps EVERY type of relationships between genes/drugs
- Advances in statistical NLP allow us to analyze every sentence

Can we characterize from text all gene-drug interaction modes, and the instances of these interactions?

CYP3A4 mRNA expression was significantly increased by rifampicin exposure in human hepatocytes.



A dependency path connects a gene to a drug in a sentence.



[*amod*, expression, *nsubjpass*, increased, *agent*, exposure, *nn*]



# Same relationships, expressed many ways

1. Atenolol, a B2AR inhibitor... appos INHIBITOR amod
2. Atenolol inhibits B2AR... nsubj INHIBITS dobj
3. Lisinopril inhibits ACE... nsubj INHIBITS dobj
4. Lisinopril is an inhibitor of ACE... nsubj INHIBITOR prep-of
5. Atenolol is a potent B2AR inhibitor... nsubj INHIBITOR amod
6. ...the B2AR inhibitor, Atenolol... appos INHIBITOR nn
7. ...Atenolol, a potent inhibitor of B2AR... appos INHIBITOR prep-of

# Same relationships, expressed many ways

1. Atenolol, a B2AR inhibitor... appos INHIBITOR amod
2. Atenolol inhibits B2AR... nsubj INHIBITS dobj
3. Lisinopril inhibits ACE... nsubj INHIBITS dobj
4. Lisinopril is an inhibitor of ACE... nsubj INHIBITOR prep-of
5. Atenolol is a potent B2AR inhibitor... nsubj INHIBITOR amod
6. ...the B2AR inhibitor, Atenolol... appos INHIBITOR nn
7. ...Atenolol, a potent inhibitor of B2AR... appos INHIBITOR prep-of

# Same relationships, expressed many ways

1. Atenolol, a B2AR inhibitor... **appos INHIBITOR amod**
2. Atenolol inhibits B2AR... **nsubj INHIBITS dobj**
3. Lisinopril inhibits ACE... **nsubj INHIBITS dobj**
4. Lisinopril is an inhibitor of ACE... **nsubj INHIBITOR prep-of**
5. Atenolol is a potent B2AR inhibitor... **nsubj INHIBITOR amod**
6. ...the ACE inhibitor, Lisinopril... **appos INHIBITOR nn**
7. ...Atenolol, a potent inhibitor of B2AR... **appos INHIBITOR prep-of**

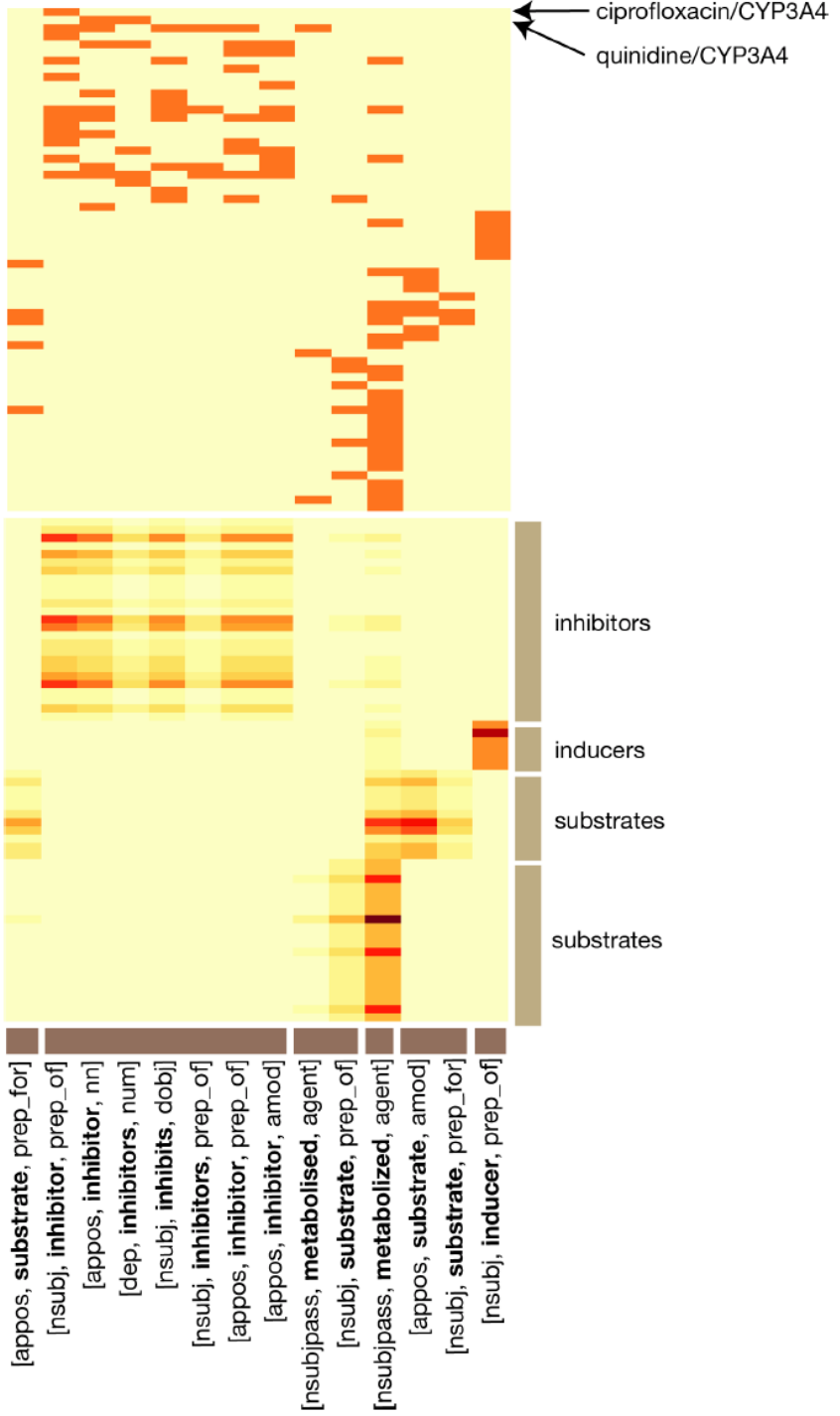
## Never observed but could imagine...

Atenolol is an inhibitor of B2AR (by analogy with 4.)

...the B2AR inhibitor, Atenolol... (by analogy with 6.)

Or

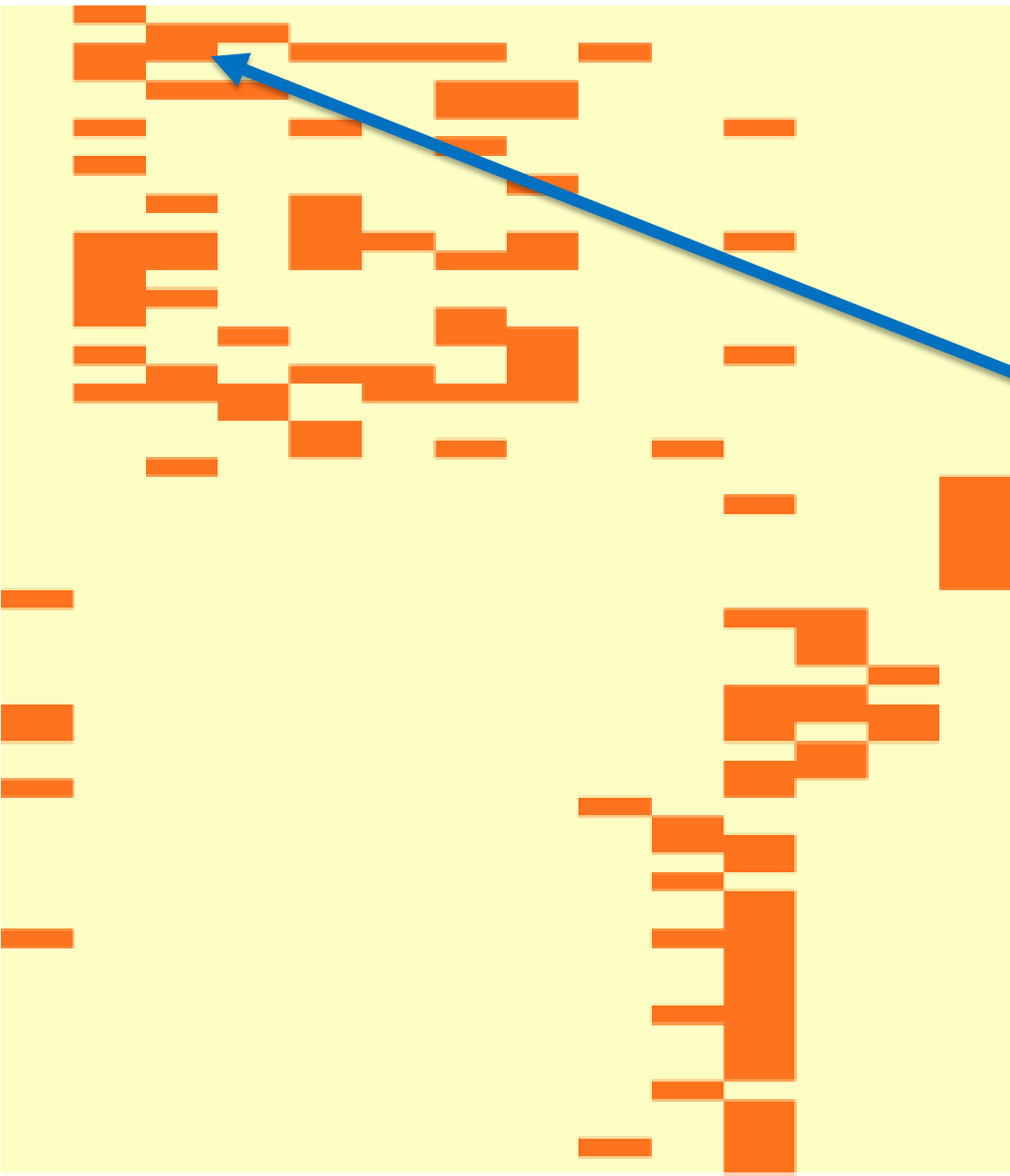
...Lisinopril, a potent inhibitor of ACE... (by analogy with 7.)



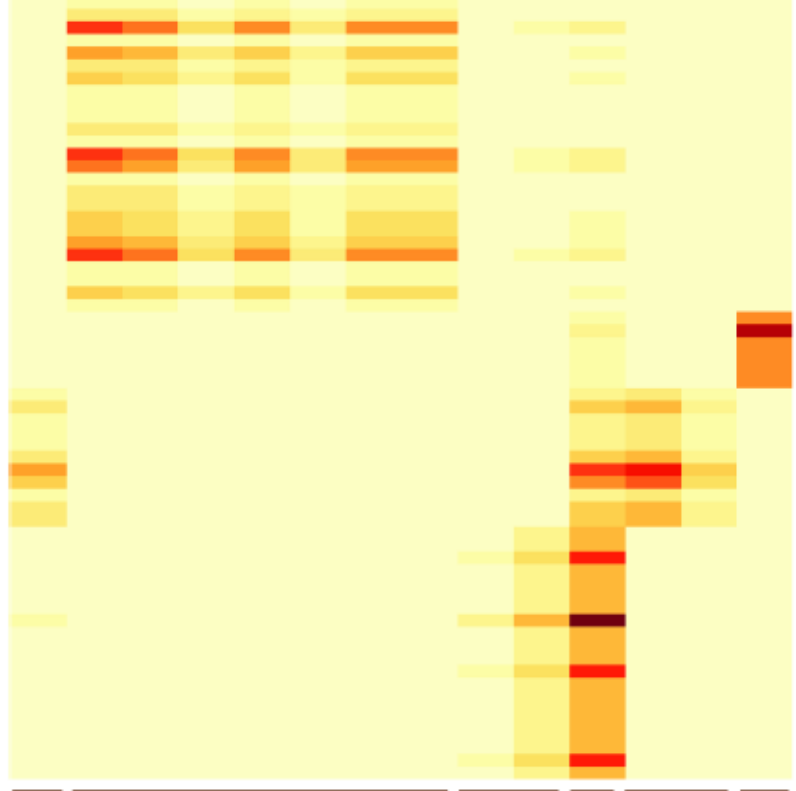
## Observed Sentences linking drug/gene

← ciprofloxacin/CYP3A4  
← quinidine/CYP3A4

Ciprofloxacin and quinidine never share a sentence but the third drug links them.



[appos, **substrate**, prep\_for]  
 [nsubj, **inhibitor**, prep\_of]  
   [appos, **inhibitor**, nn]  
   [dep, **inhibitors**, num]  
     [nsubj, **inhibits**, dobj]]  
 [nsubj, **inhibitors**, prep\_of]  
 [appos, **inhibitor**, prep\_of]  
   [appos, **inhibitor**, amod]  
 [nsubjpass, **metabolised**, agent]  
   [nsubj, **substrate**, prep\_of]  
 [nsubjpass, **metabolized**, agent]  
   [appos, **substrate**, amod]  
 [nsubj, **substrate**, prep\_for]  
   [nsubj, **inducer**, prep\_of]



substrates

substrates

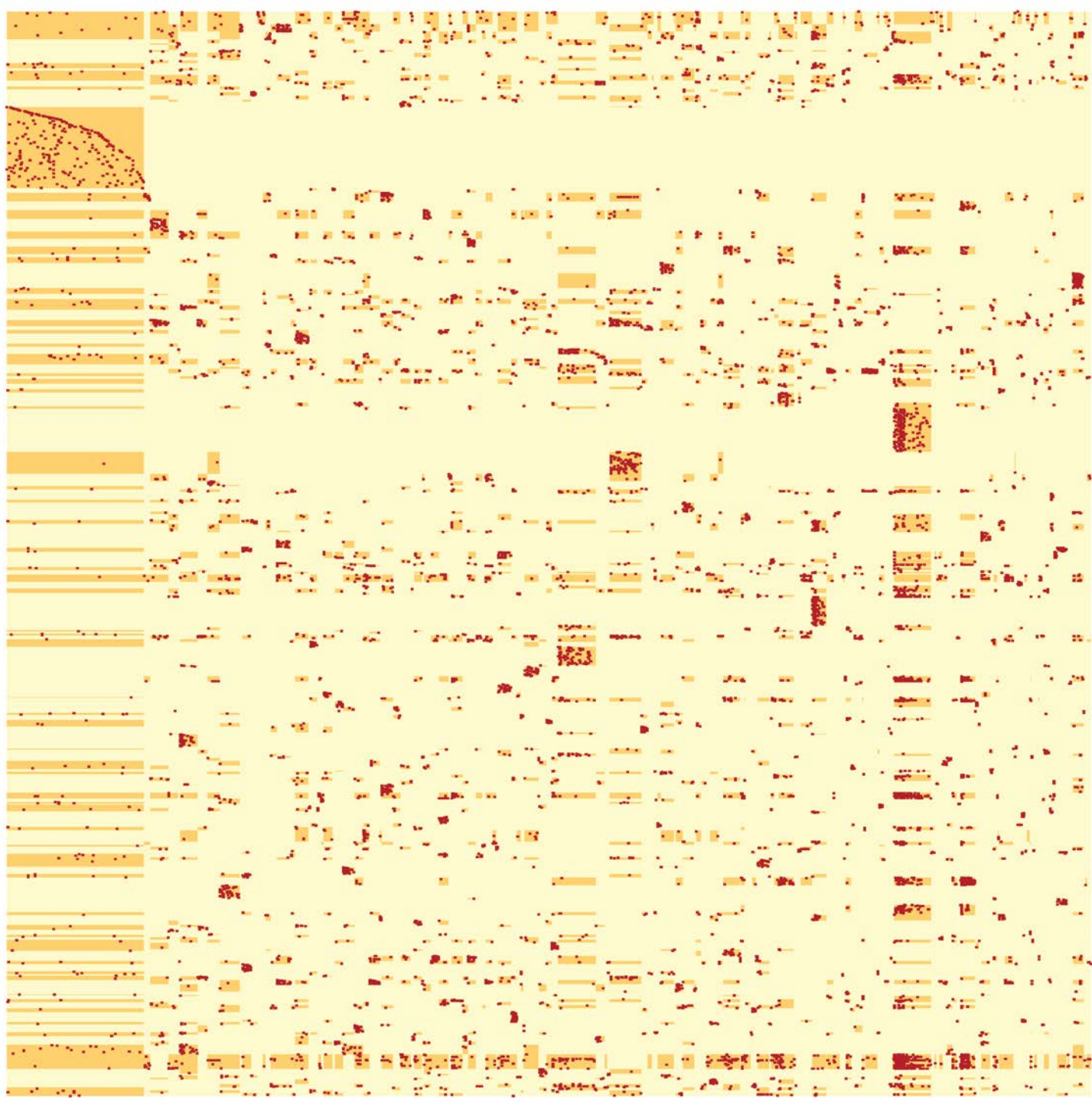
inducers

inhibitors

**PubMed  
gene/drug  
space  
bicluster**

**Rows =  
~50K  
drug/gene  
pairs**

**Cols =  
197K  
dependency  
paths**



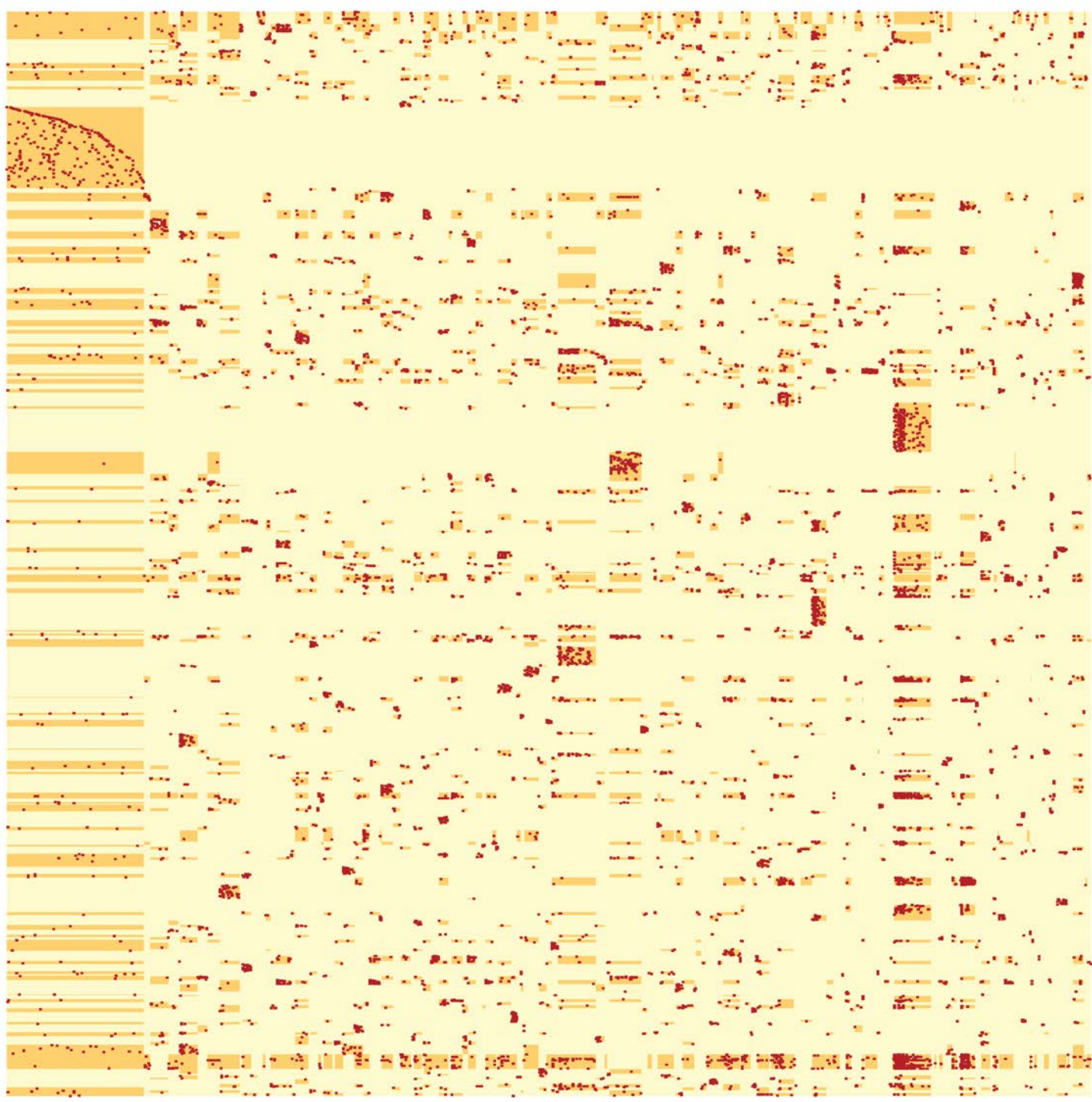
	Dependency path	Example sentence (PubMed ID)	Frequency
[1]	[ <i>appos</i> , inhibitor, <i>amod</i> ]	<b>Geldanamycin</b> (GA), an <b>HSP90</b> inhibitor, is able to suppress 1,25-induced differentiation of HL60 cells. (20138989)	1181
[2]	[ <i>appos</i> , inhibitor, <i>prep_of</i> ]	The mNQO activity was insensitive to <b>dicoumarol</b> , a potent inhibitor of cytosolic <b>NQO1</b> . (10683249)	452
[3]	[ <i>appos</i> , antagonist, <i>amod</i> ]	The recommended therapy for stage III disease, based on clinical trials and by the Israeli Ministry of Health for 2006, includes <b>bosentan</b> (Tracleer), an <b>endothelin-1</b> antagonist. (18686806)	338
[4]	[ <i>nsubjpass</i> , metabolized, <i>agent</i> ]	<b>Amodiaquine</b> is mainly metabolized hepatically towards its major active metabolite desethylamodiaquine, by the polymorphic P450 isoform <b>CYP2C8</b> . (18855526)	204
[5]	[ <i>nsubj</i> , inhibits, <i>dobj</i> ]	<b>Salbutamol</b> inhibits <b>IFN-gamma</b> and enhances IL-13 production by PBMCs from asthmatics. (20523061)	118



**PubMed  
gene/drug  
space  
bicluster**

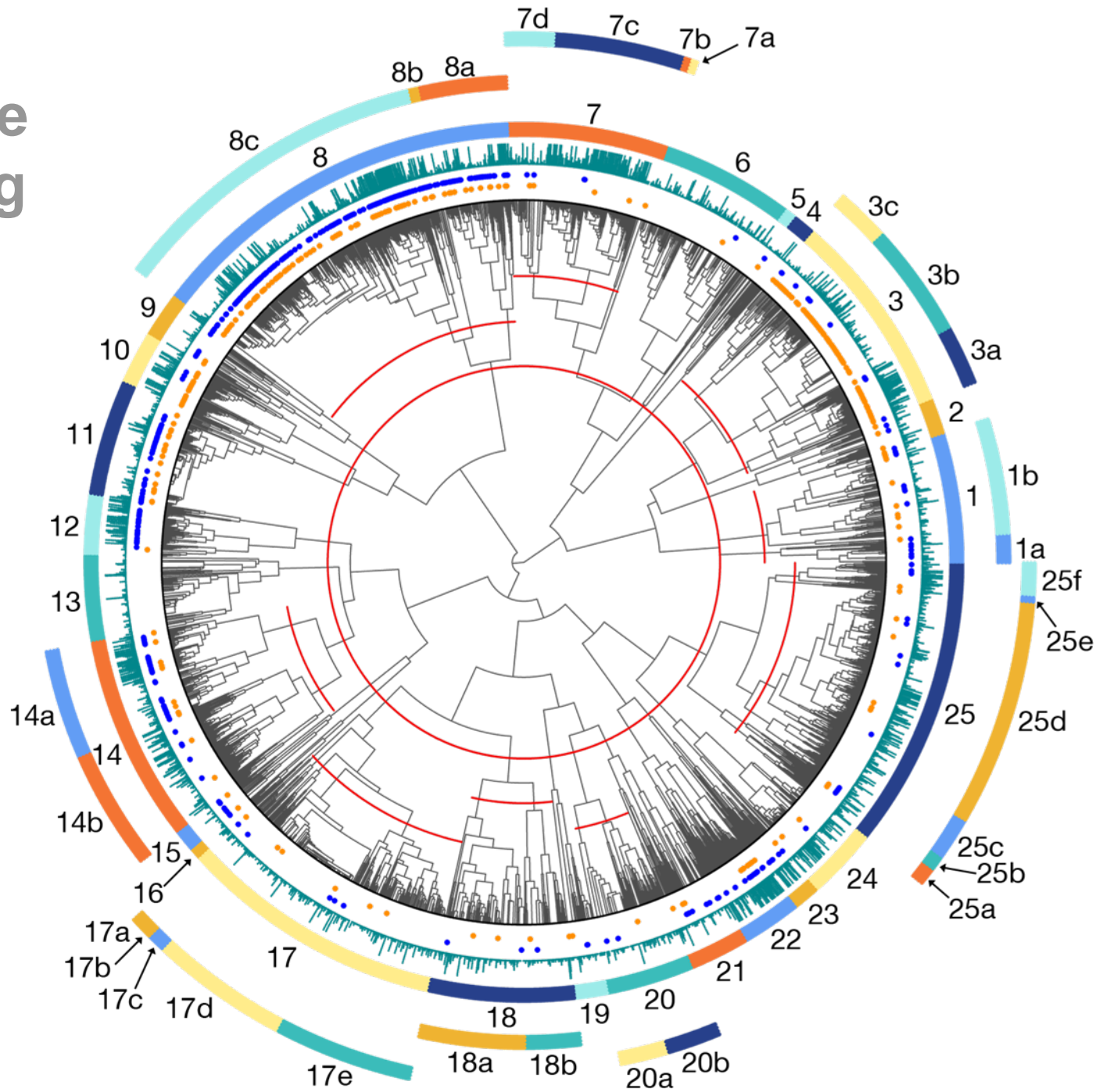
**Rows =  
~50K  
drug/gene  
pairs**

**Cols =  
197K  
dependency  
paths**

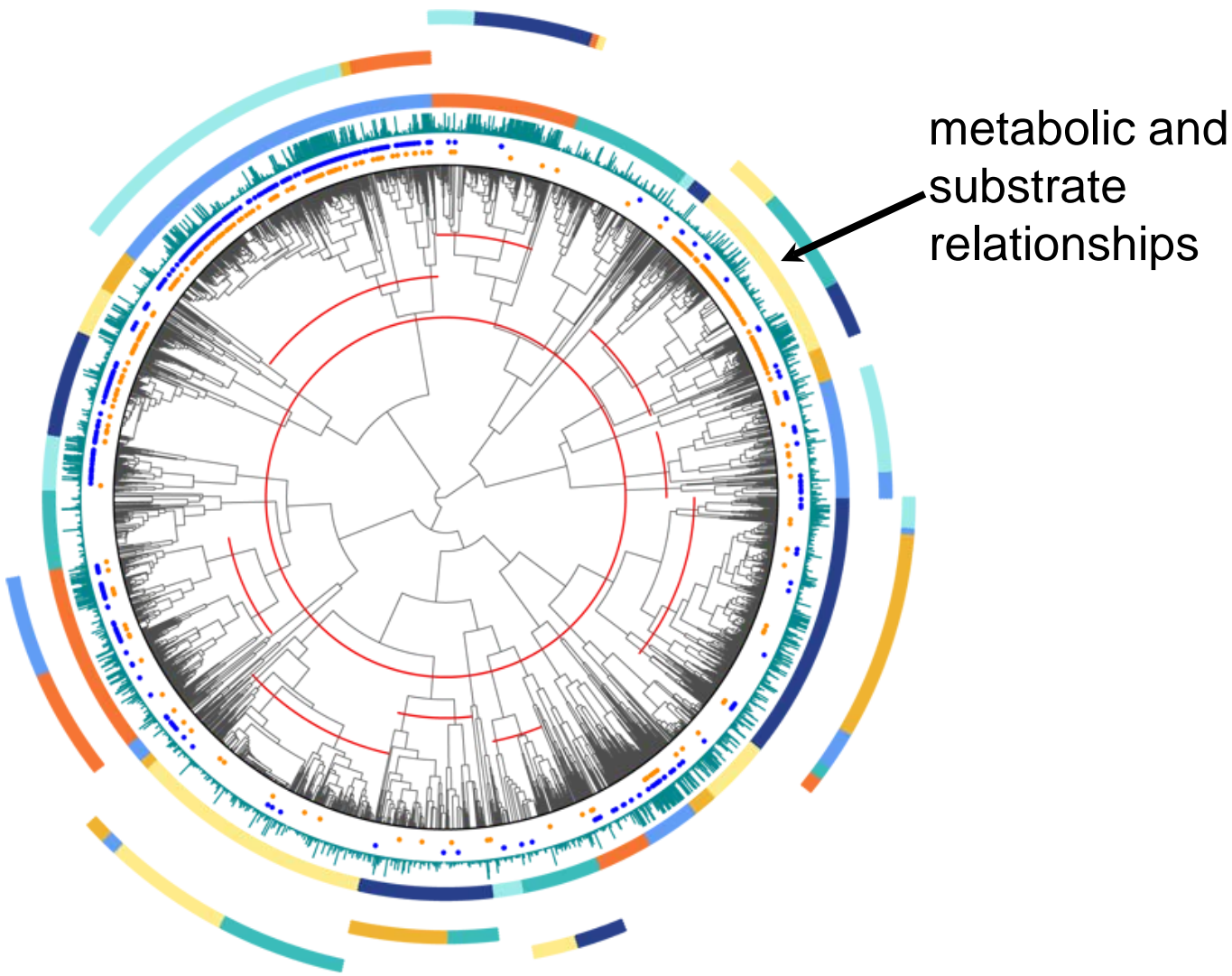


First Pattern	Second Pattern	Frequency of co-clustering
<p>[<i>nsubj</i>, antibody, <i>partmod</i>, directed, <i>prep_against</i>]  <i>D</i> is an antibody directed against <i>G</i>.</p>	<p>[<i>nsubj</i>, antibody, <i>partmod</i>, targeting, <i>dobj</i>]  <i>D</i> is an antibody targeting <i>G</i>.</p>	0.59
<p>[<i>prep_such_as</i>, inhibitor, <i>amod</i>]  <i>G</i> inhibitor such as <i>D</i></p>	<p>[<i>prep_including</i>, inhibitors, <i>amod</i>]  <i>G</i> inhibitors, including <i>D</i></p>	0.31
<p>[<i>prep_such_as</i>, agonists, <i>nn</i>]  <i>G</i> agonists, such as <i>D</i>, ...</p>	<p>[<i>amod</i>, activators, <i>nn</i>]  <i>G</i> activators, <i>D</i> and...</p>	0.12
<p>[<i>nsubjpass</i>, metabolized, <i>agent</i>]  <i>D</i> is metabolized by <i>G</i></p>	<p>[<i>dep</i>, substrates, <i>nn</i>]  <i>G</i> substrates (<i>D</i>, ...), ...</p>	0.11
<p>[<i>nsubj</i>, blocked, <i>dobj</i>, activation, <i>amod</i>]  <i>D</i> blocked <i>G</i> activation</p>	<p>[<i>nsubj</i>, inhibited, <i>dobj</i>]  <i>D</i> inhibited <i>G</i></p>	0.07
<p>[<i>nsubj</i>, increased, <i>dobj</i>, expression, <i>prep_of</i>, mRNA, <i>nn</i>]  <i>D</i> increased the expression of <i>G</i> mRNA</p>	<p>[<i>nsubj</i>, induces, <i>dobj</i>, activity, <i>amod</i>]  <i>D</i> induces <i>G</i> activity</p>	0.03

# The universe of gene-drug relationship types



# Metabolic relationships form a distinct cluster enriched for drug-gene pairs known to PharmGKB

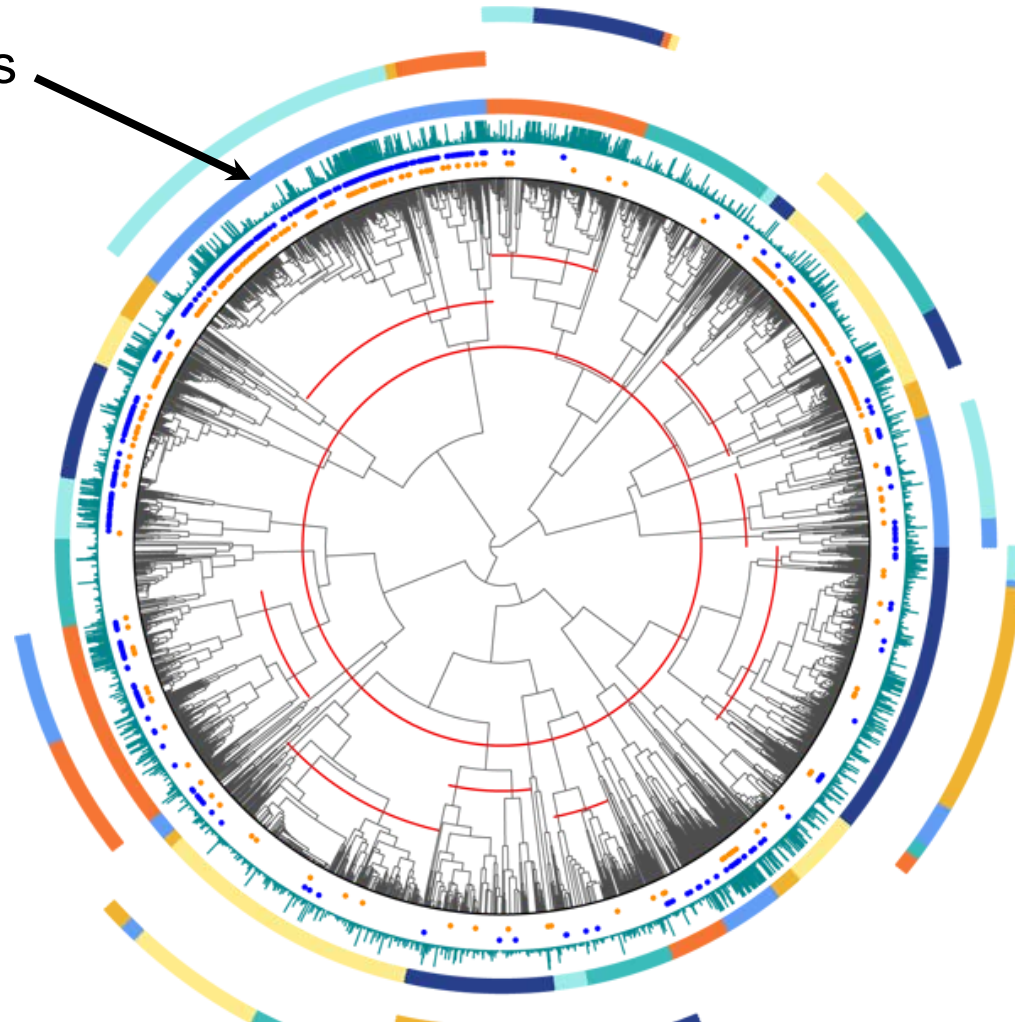


metabolic and substrate relationships

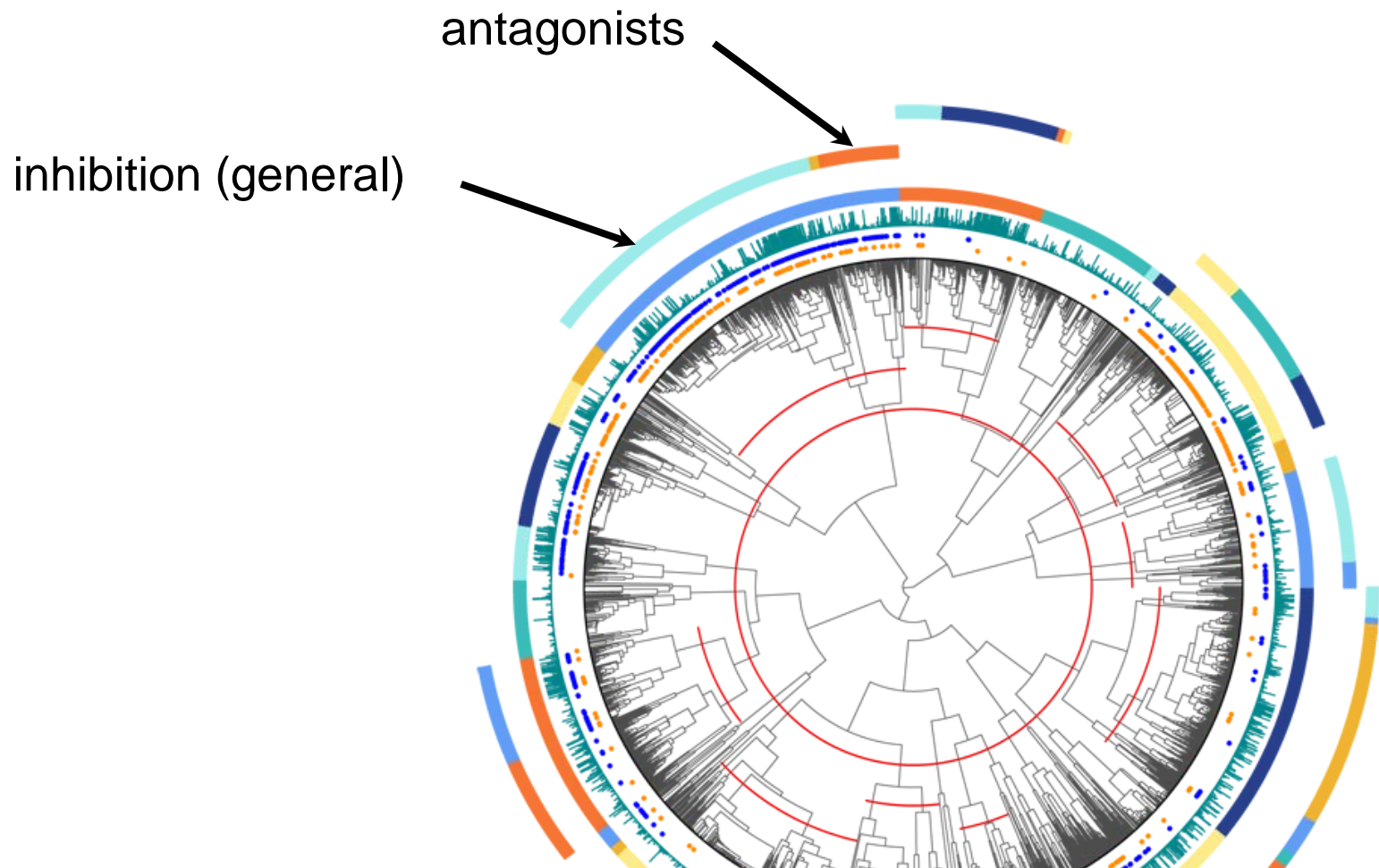


# Inhibitory relationships also form a cluster enriched for both PharmGKB and DrugBank relations

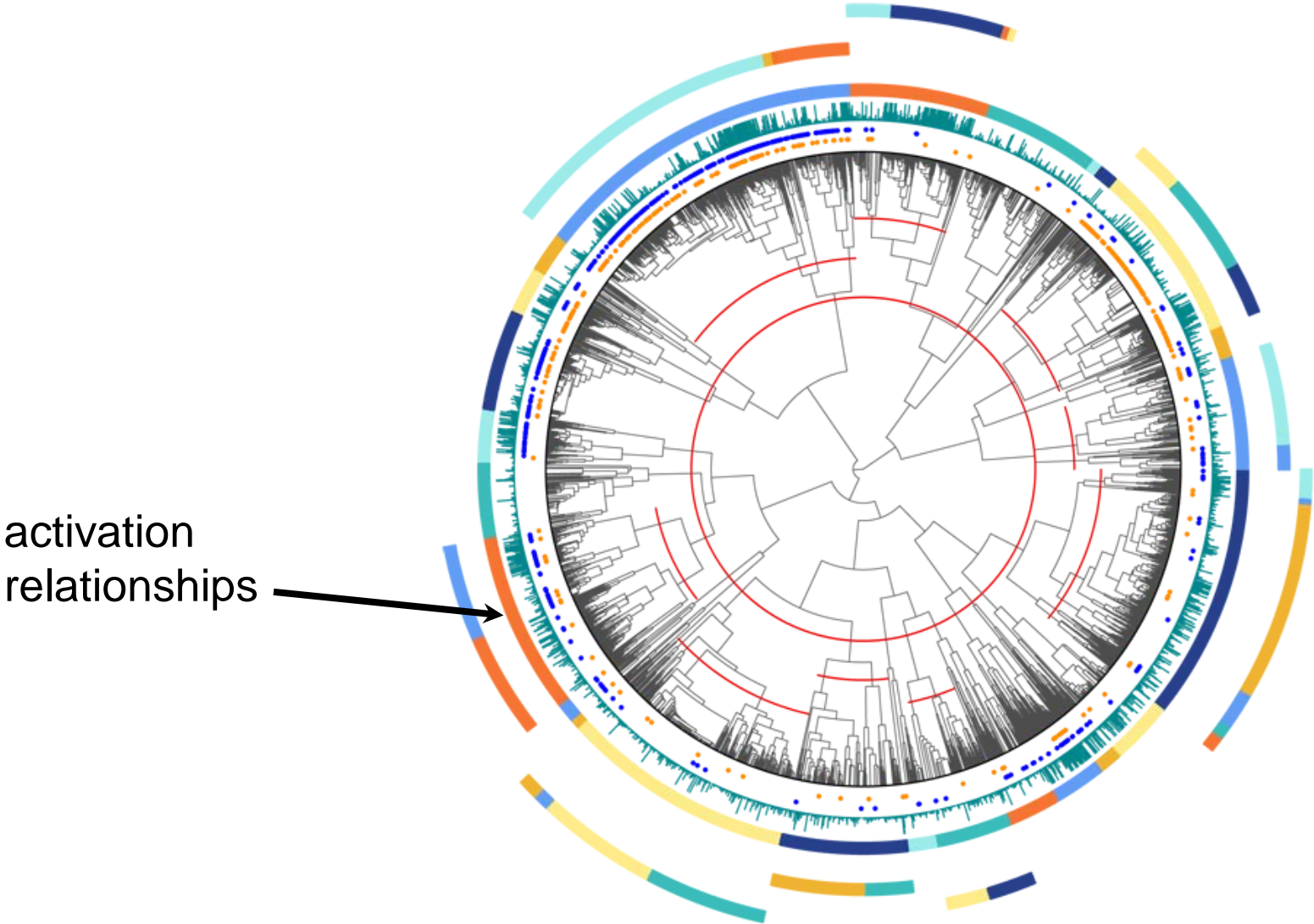
inhibitory relationships



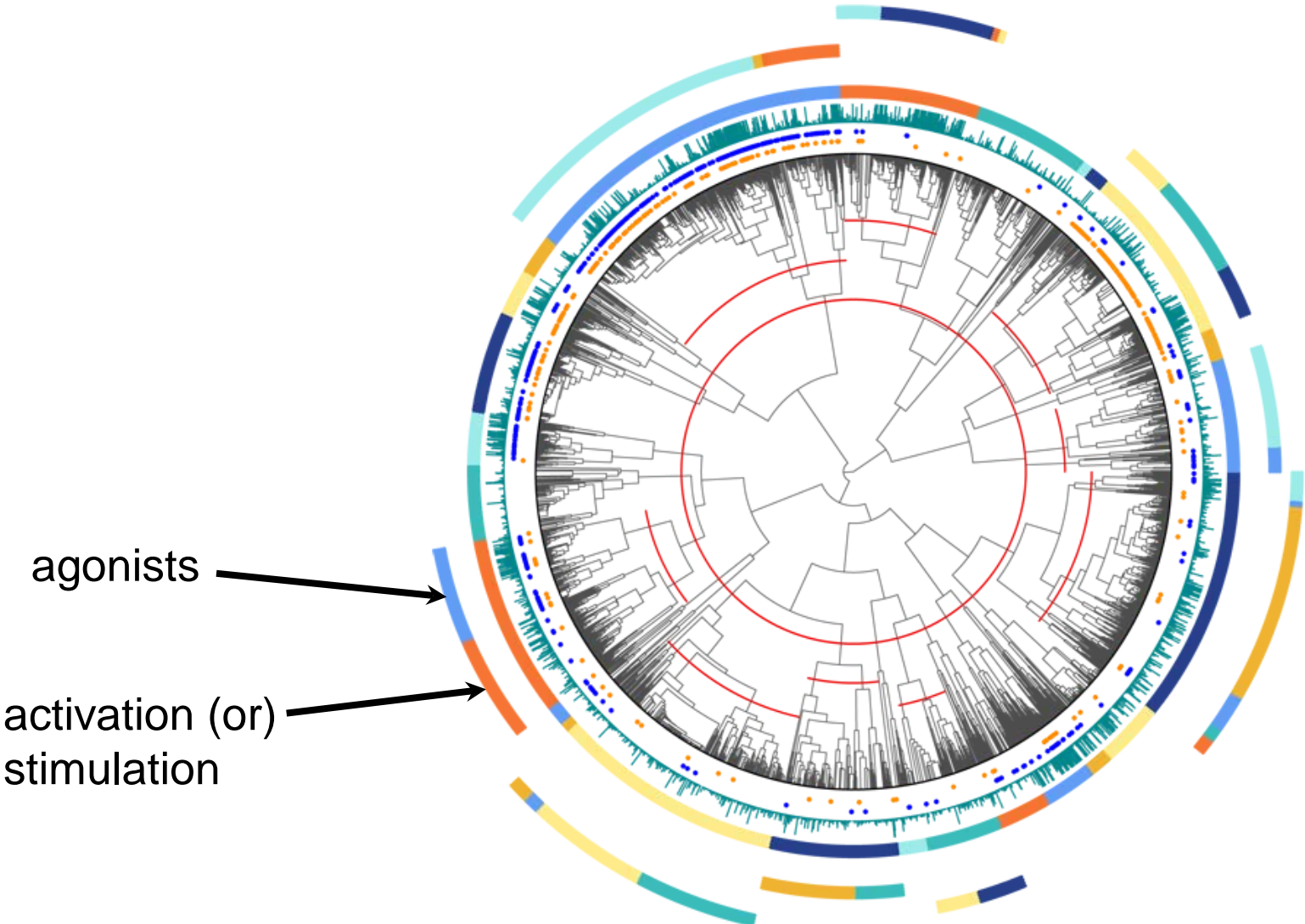
# The algorithm learns that antagonism is a subset of inhibition



# Activation relationships also form a distinct cluster



# The algorithm learns that agonism is a subset of activation





# Gene-Drug interaction types

- G synthesizes D
- G activates D
- G enzymatically modifies D
- D substrate of G
- G metabolizes D
- D substrate of G
- D indirectly effects G
- D coadministered w/ G
- D increases levels of G
- D raises levels of G
- D antagonizes G
- D inhibits G
- D interacts with G
- D binds/inhibits G
- G receptor for D
- D activates G
- D agonist for G
- D binds G
- D induces expression G
- D affects expression G
- D suppresses expression G
- D inhibits activation of G
- D subtly inhibits G

**Table 5. Top 20 predictions of new drug-gene relationships for PharmGKB, and whether a PGx relationship has been documented in the literature.**

	Candidate drug-gene pair	Relative certainty	Literature reference (PMID)	Comment
[1]	omeprazole, CYP2C19	1.000	11069321	*** Individual polymorphisms of CYP2C19 already associated with omeprazole in PharmGKB.
[2]	mexiletine, CYP1A2	0.995	9690950	**
[3]	fentanyl, P-gp	0.994	17192767	***
[4]	voriconazole, CYP3A4	0.986	17433262	**
[5]	cyclosporine, CYP3A4	0.983	18978522	*** Association listed in PharmGKB as “ambiguous”.
[6]	duloxetine, CYP1A2	0.983	18307373	**
[7]	fluconazole, UGT2B7	0.982	16542204	**
[8]	montelukast, CYP2C8	0.973	21838784	**
[9]	dydrogesterone, AKR1C1	0.968	20727920	**
[10]	voriconazole, CYP2C9	0.966	16940139	*
[11]	imipramine, FMO1	0.962	19262426	*** Experiment conducted in mice.
[12]	ticlopidine, CYP2C19	0.961	21178986	*
[13]	moclobemide, MAO-B	0.960	7586937	In this article, MAO-B activity was studied in relation to moclobemide response, but specific polymorphisms were not investigated.
[14]	ritonavir, P-gp	0.958	16184031	*** Association listed in PharmGKB as “ambiguous”.
[15]	cyclosporin, MDR1	0.955	15116055	*
[16]	cyclosporin, P-gp	0.952	15116055	* Same gene as 15.
[17]	vinblastine, P-gp	0.951	16917872	*** Association listed in PharmGKB as “ambiguous”.
[18]	amprenavir, CYP3A4	0.950	9649346	**
[19]	perazine, CYP1A2	0.945	11026737	**
[20]	lopinavir, ABCB1	0.939	21743379	*

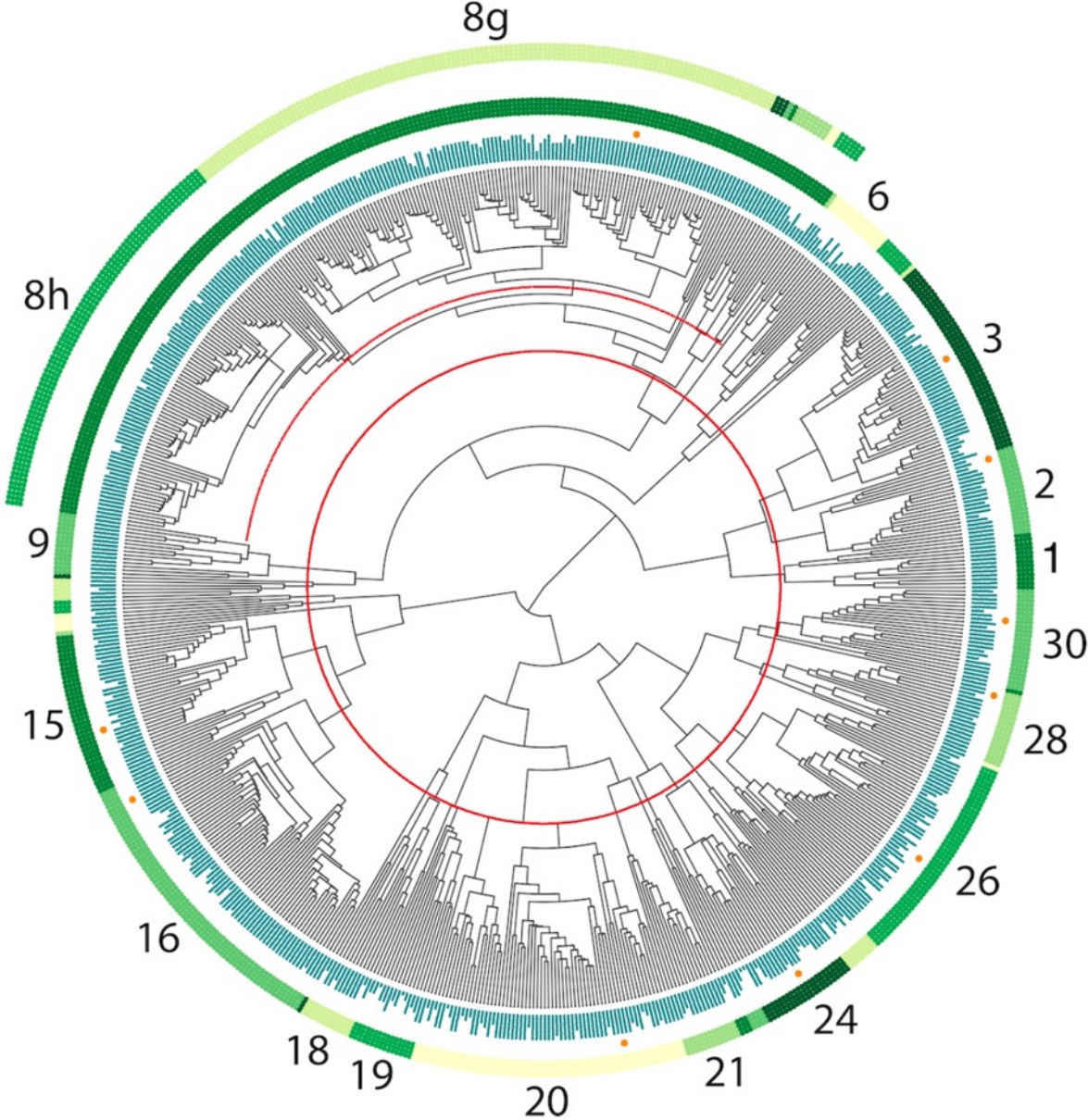
\*\*\* indicates that an association has been demonstrated experimentally between changes in the expression/activity of the gene/protein and the efficacy of the drug

\*\* indicates that such an association is likely, but has not yet been studied

\* indicates that the association has been studied experimentally, and the experiment refuted the association. Here we include only associations between pharmaceutical compounds and single genes; predicted associations involving endogenous compounds and/or groups of genes are included in the supplement, however.

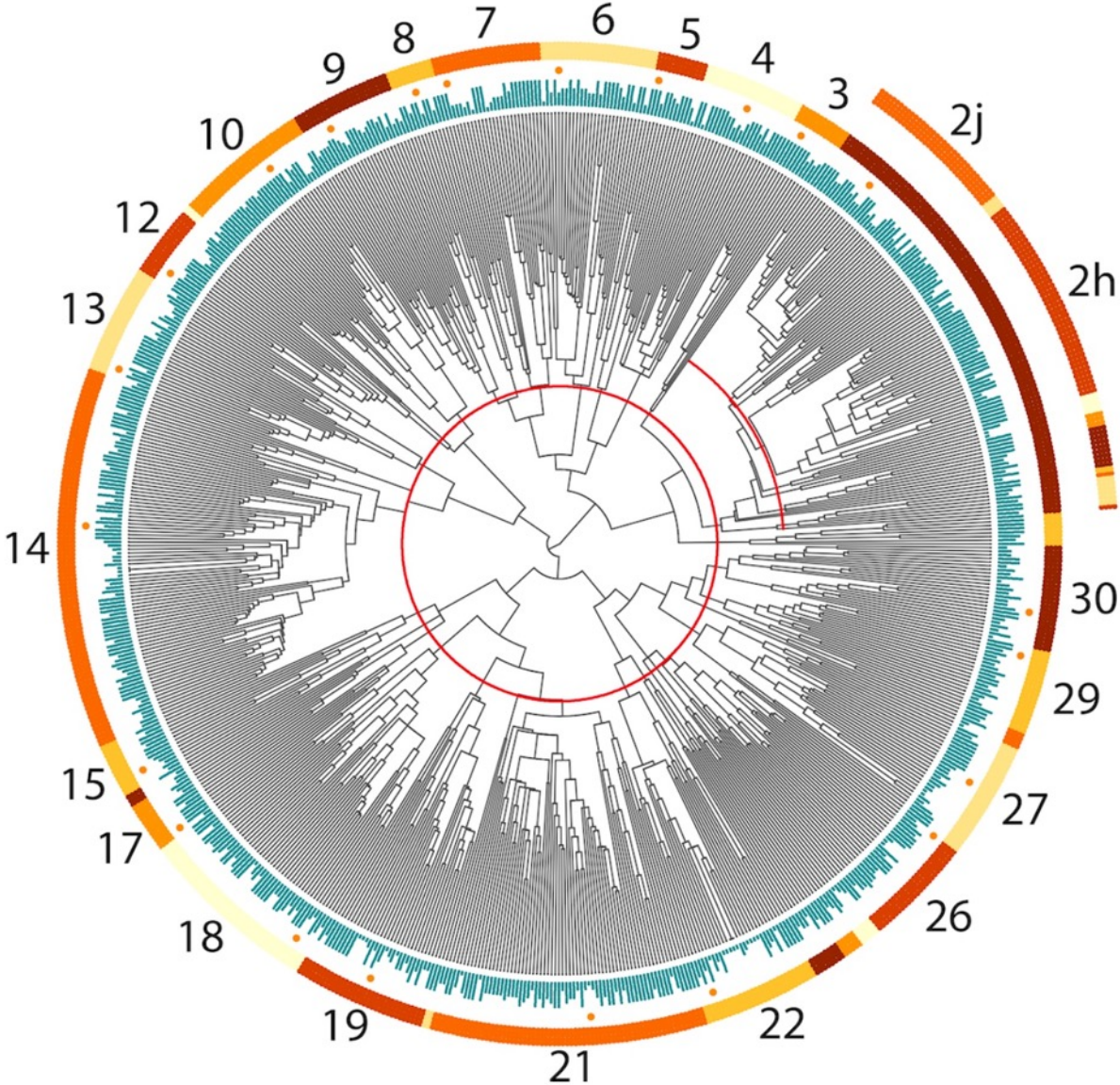
	gene pair	certainty	reference (PMID)		
[1]	omeprazole, CYP2C19	1.000	11069321	***	Individual polymorphisms of CYP2C19 in PharmGKB.
[2]	mexiletine, CYP1A2	0.995	9690950	**	
[3]	fentanyl, P-gp	0.994	17192767	***	
[4]	voriconazole, CYP3A4	0.986	17433262	**	
[5]	cyclosporine, CYP3A4	0.983	18978522	***	Association listed in PharmGKB as "an"
[6]	duloxetine, CYP1A2	0.983	18307373	**	
[7]	fluconazole, UGT2B7	0.982	16542204	**	
[8]	montelukast, CYP2C8	0.973	21838784	**	
[9]	dydrogesterone, AKR1C1	0.968	20727920	**	
[10]	voriconazole, CYP2C9	0.966	16940139	*	
[11]	imipramine, FMO1	0.962	19262426	***	Experiment conducted in mice.
[12]	ticlopidine, CYP2C19	0.961	21178986	*	
[13]	moclobemide, MAO-B	0.960	7586937		In this article, MAO-B activity was studied in response, but specific polymorphisms
[14]	ritonavir, P-gp	0.958	16184031	***	Association listed in PharmGKB as "an"
[15]	cyclosporin, MDR1	0.955	15116055	*	
[16]	cyclosporin, P-gp	0.952	15116055	*	Same gene as 15.
[17]	vinblastine, P-gp	0.951	16917872	***	Association listed in PharmGKB as "an"
[18]	amprenavir, CYP3A4	0.950	9649346	**	
[19]	perazine, CYP1A2	0.945	11026737	**	

# Chemical-Disease relationships

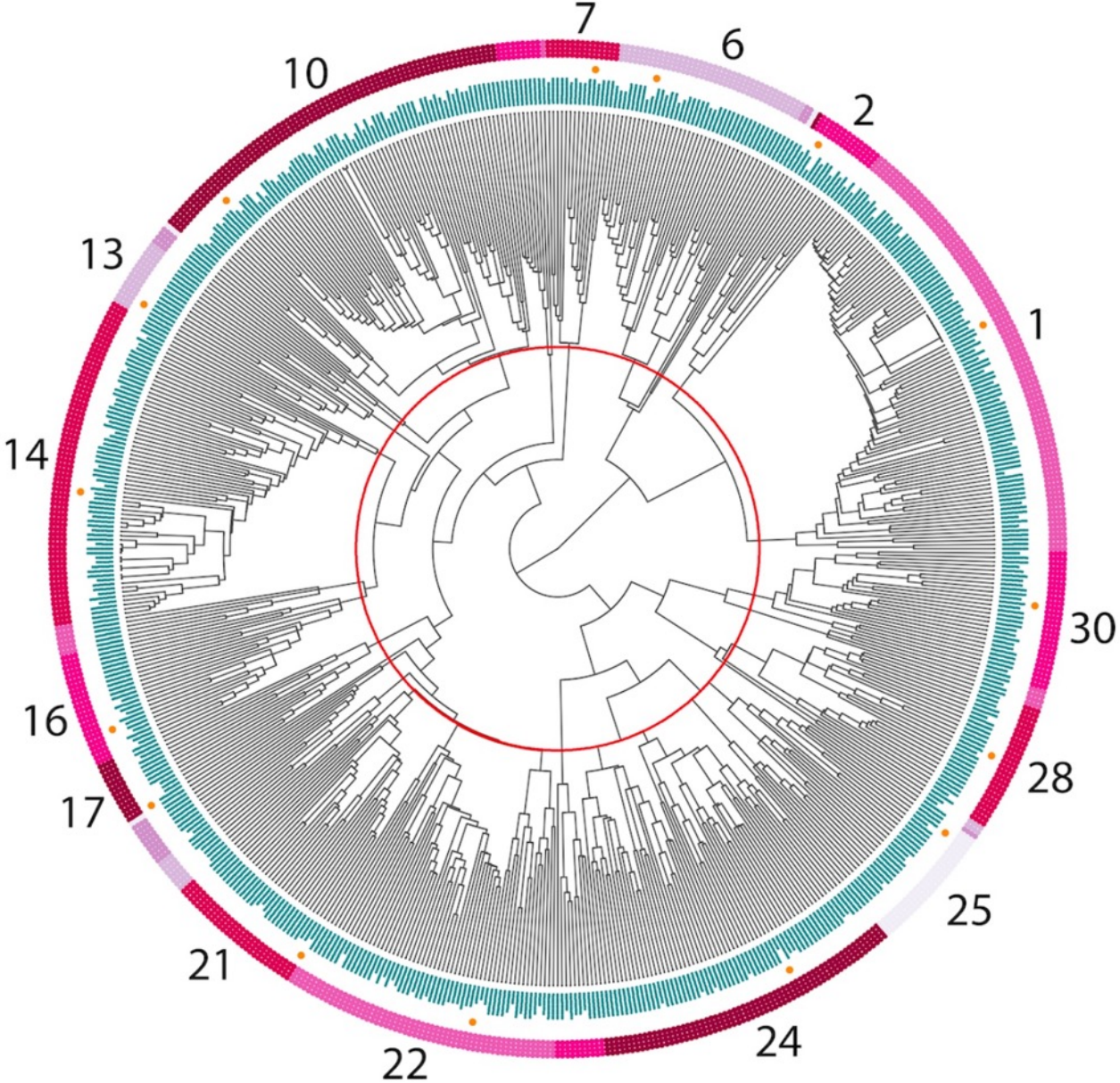




# Gene-Disease relationships



# Gene-Gene relationships



# Large scale extraction of PGx knowledge from full text

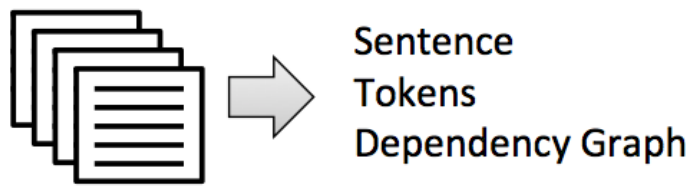
- Increasingly available full text
- Availability of technology for “trained systems” to find entities and relationships in
- DeepDive developed by Chris Re

Can we recognize key PharmGKB entities and their relationships from full text automatically?

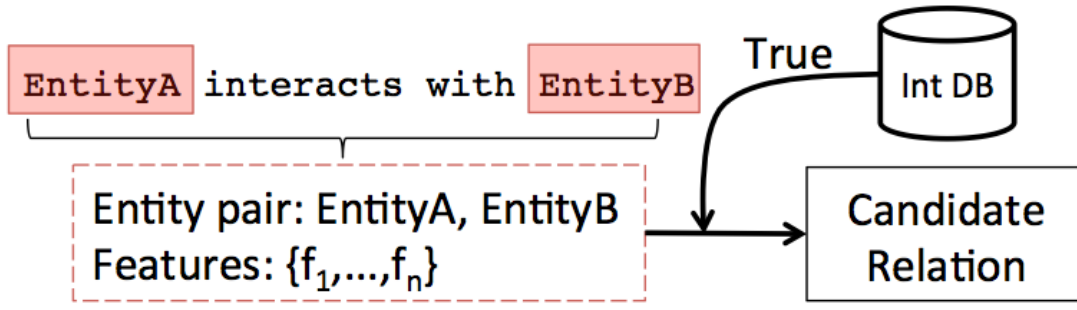


# Deep Dive Pipeline for extracting Gene-Gene interactions

## A. Text Preprocessing



## B. Relation Extractor

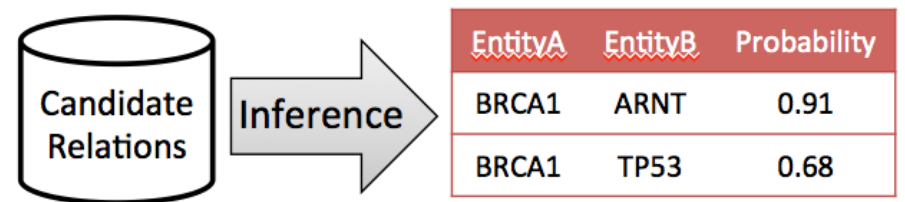


## D. System Tuning

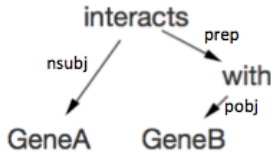
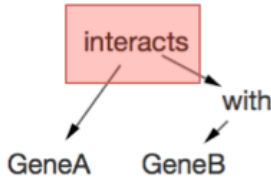
EntityA	EntityB	Probability	Interaction?
BRCA1	ARNT	0.91	YES
BRCA1	TP53	0.68	<b>YES</b>

- Fix errors
- Snowball

## C. DeepDive





Feature Type	Example	System Representation
Dependency Path	 <pre> graph TD     interacts -- nsubj --&gt; GeneA     interacts -- prep --&gt; with     with -- pobj --&gt; GeneB </pre>	<pre>nsubj(interacts, GeneA), prep(interacts, with), pobj(with, GeneB)</pre>
Prepositional Pattern	<p>Binding of GeneA and GeneB.</p>	<pre>prep_pattern_[Binding_of]</pre>
Verb on Dependency Path	 <pre> graph TD     interacts -- nsubj --&gt; GeneA     interacts -- prep --&gt; with     with -- pobj --&gt; GeneB </pre>	<pre>verb_on dep_path_[interact]</pre>
1-Word Window	<p>GeneA forms a complex with GeneB in vitro.</p>	<pre>g1_right_1gram_[form]</pre>
2-Word Window	<p>GeneA forms a complex with GeneB in vitro.</p>	<pre>g1_right_2gram_[form_a]</pre>
Word Sequence Window	<p>GeneA regulates GeneB in humans.</p>	<pre>word_seq_[regulate]</pre>

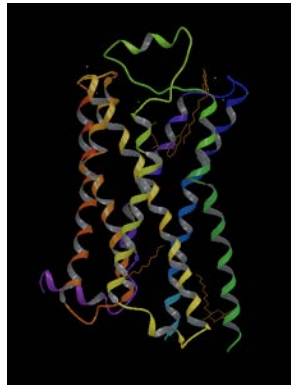
**Table 1.** Top 10 positive gene–gene features from DeepDive

---

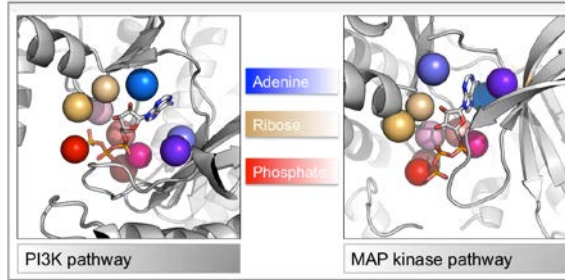
Feature	Weight
Single_Verb_Between_Genes_[bind]	1.25
Single_Verb_Between_Genes_[interact]	1.07
Verb_On_Dependency_Path_[bind]	0.91
Verb_On_Dependency_Path_[interact]	0.74
Single_Verb_Between_Genes_[regulate]	0.67
Verb_Between_Genes_[bind]	0.63
Verb_On_Dependency_Path_[regulate]	0.58
Window_Left_Gene1_Phrase_[GENE and]	0.57
Window_Right_Gene2_1gram_[protein]	0.57
Window_Left_Gene1_Phrase_[interaction between]	0.51

---

# Thus, the emerging network for drugs....

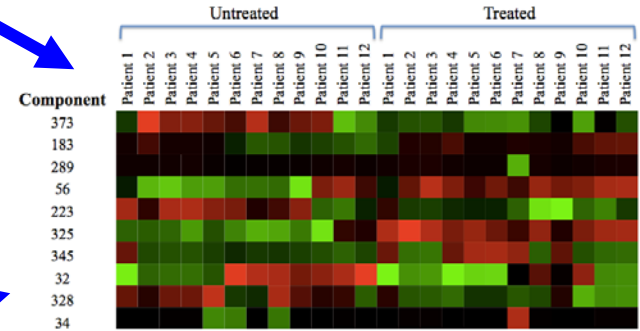


Target structure & dynamics

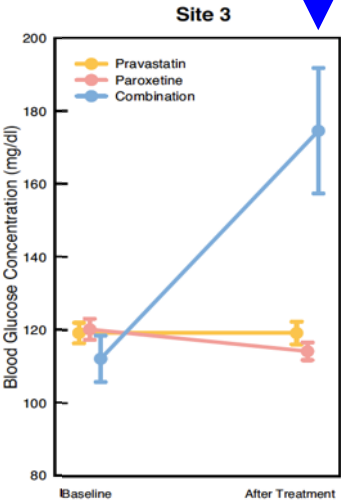
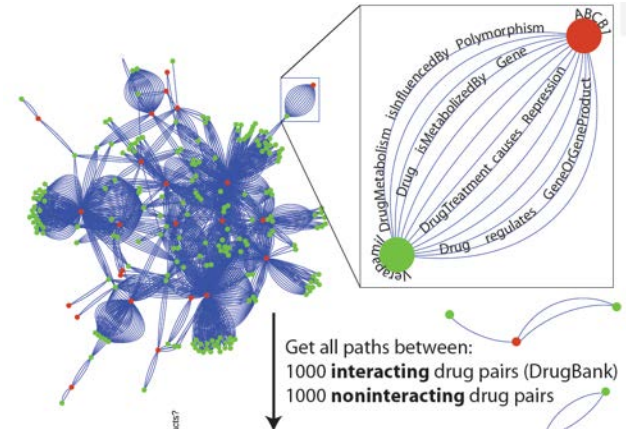


Drug recognition & binding

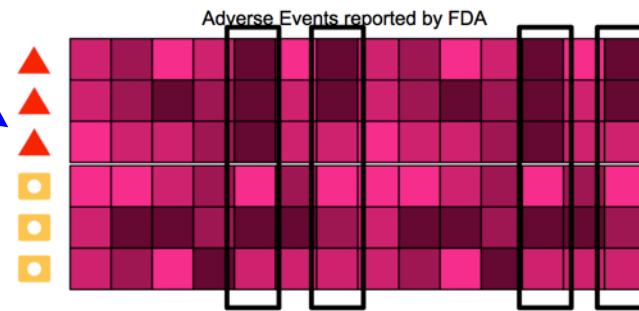
Cellular response & pathways



Text mining of gene, drug, phenotype associations



Clinical response datamining



Population effect reporting



Beth Percha, BMI



Emily Mallory, BMI

Support:

NIH GM61374

NIH LM05652

NIH GM102365

NIH MH094267

NIH HL117798

Oracle

Pfizer

Microsoft

FDA



Yuhao Zhang, BMI



Chris Re

# Thanks to PharmGKB Team

Teri Klein

Michelle Carrillo

Team: Maria Avarelllos, Julia Barbarino,  
Lester Carter, Matt Devlin, Alie Fohner, Li  
Gong, Tiffany Murray, Katrin Sangkuhl,  
Caroline Thorn, Ryan Whaley, Mark Woon

NIH: GM-61374



# Radio Show on SiriusXM 121

“The Future of Everything” w/ Russ

Joint effort with Stanford University to discuss science and technology and the future.

Saturdays at 8 AM ET, freely available at:

[stanfordradio.stanford.edu](http://stanfordradio.stanford.edu)

